Geraldo A. Passos *Editor*

Transcriptomics in Health and Disease

Second Edition

Springer

Transcriptomics in Health and Disease

Geraldo A. Passos
Editors

Transcriptomics in Health and Disease

Second Edition

Springer

Editor
Geraldo A. Passos
Molecular Immunogenetics Group, Department of Genetics
School of Medicine of Ribeirão Preto, University of São Paulo
Ribeirão Preto, Brazil

Laboratory of Genetics and Molecular Biology
Department of Basic Oral Biology
School of Dentistry of Ribeirão Preto
University of São Paulo
Ribeirão Preto, Brazil

ISBN 978-3-030-87820-7 ISBN 978-3-030-87821-4 (eBook)
https://doi.org/10.1007/978-3-030-87821-4

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book is dedicated to the Molecular Immunogenetics Group, Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Brazil, which in 2021 celebrates its 25 years of research, teaching, and advising graduate students.

Preface

The completion of the human genome, with its more than three billion base pairs (bp) of sequenced DNA, has provided an unprecedented wealth of knowledge. With the additional investigation of single nucleotide polymorphisms (SNPs), we have also learned how little genetic variability there truly is in the human genome. Moreover, genome-wide association studies (GWAS) have revealed critical genotype-phenotype correlations. Nevertheless, our understanding of the functionality of the genome is still in its beginnings.

Dispersed among its three billion bp, the human genome features approximately 20–25,000 functional genes that encode a vast set of proteins and their isoforms, which compose the human proteome. In recent years, however, scientists realized that the functionality of the genome is not restricted to only protein-encoding genes, which are transcribed into messenger RNAs, but also to the transcription of non-coding RNAs [e.g., microRNAs (miRNAs) and long non-coding RNAs (lncRNAs)], which play essential roles in the posttranscriptional control of gene expression and, consequently, influence the resulting phenotypes.

At this point – from studies investigating where the functions of the genome first begin – the science of transcriptomics emerged. For example, how are RNA molecules transcribed, what are the different RNA species, what are the functions of each species, what are the different mRNA isoforms, and how are they differentially expressed among cells, tissues, and organs?

Transcriptomics can therefore be thought of as the molecular biology of gene expression on a large scale. It is derived from functional genomics studies with a focus on transcription. Since its origin, transcriptomics has benefitted from and will continue to benefit from microarray technology. The RNA sequencing (RNA-Seq) is undoubtedly the ultimate tool for delving into the differences at the sequence level or confirming the specific RNA isoform involved. Moreover, more and more questions from current projects are these. Even more so now, with the emergence of new technologies for high-throughput RNA-Seq, we can answer more questions about the structure of RNAs, such as those found in alternative splicing. However, the bottleneck remains in the data analysis because sequences are currently being obtained in quantities that have never been previously achieved.

However, as microarray bioinformatics has reached a very advanced stage (with more than 20 years to perfect the analysis pipeline) and as microarray slides themselves have become increasingly “large,” currently encompassing sequences from the entire functional genome plus the complete set of known non-coding RNAs, researchers have not neglected the applications of this important technology.

Recent comparative analyses have indicated a strong concordance between exon microarrays and RNA-Seq data. Therefore, the goal is now to use these two complementary strategies for in-depth transcriptomics studies.

The advances of the last 5 years have made possible the sequencing of the single-cell transcriptome and showed us how individual cells respond to normal and pathological differentiation stimuli. More recently, space transcriptome technology has emerged, making it possible to investigate how gene expression varies in specific locations in a tissue, organ, or cancerous tumor.

This book was organized based on these assumptions. It includes 18 chapters and covers the fundamental concepts of transcriptomics and the current analytical methods. We provide high-level technical and scientific examples, using accessible language whenever possible, as each chapter is written by experienced and productive researchers in the field.

Over the first 10 chapters (Part I), we introduce the concept of the transcriptome, the alternative processing of pre-mRNAs, as well as how microarrays or RNA-Seq can be used to trace expression signatures, measure transcriptional expression levels, and establish connections between genes based on their transcriptional activity in normal cells, differentiating cells, and organs. Moreover, we introduce dedicated chapters on proteomics, single-cell RNA-Seq, and spatial transcriptomics as well.

Chapters 11, 12, 13, 14, 15, 16, 17, and 18 (Part II) then provide examples of the state of the transcriptome associated with major human diseases, such as autoimmune diseases, metabolic diseases (such as type 2 diabetes mellitus), genetic diseases, cancer, and infections caused by pathogenic microorganisms, such as fungi or the protozoan *Trypanosoma cruzi*, which is the causative agent of Chagas disease.

I hope this book will be helpful to researchers who wish to gain a comprehensive view of transcriptomics in health and human disease. I want to thank all of the authors for their dedication and time spent writing these chapters. Finally, I thank Springer Nature for providing this opportunity and their continued support during the writing and organization of this work.

Ribeirão Preto, São Paulo, Brazil Geraldo A. Passos

Contents

Part I Basic Principles of the Transcriptome and Its Analysis

Part II Transcriptome in Disease

About the Editor

Geraldo A. Passos received his PhD degree in biochemistry (1988) from Ribeirão Preto Medical School, University of São Paulo (USP), Brazil. His postdoctoral studies were conducted at the Molecular Genetics Institute of Montpellier (CNRS), France (1992–1994) with sequencing and physical mapping of the human immunoglobulin lambda locus on chromosome 22q11.2. For several years, he worked in close collaboration with the microarray laboratory at Centre d'Immunologie de Marseille-Luminy (CIML, Marseille, France) (1999–2001) and then the Unité INSERM 1090 in Marseille (2002–2017) to study the large-scale gene expression in the thymus. He is currently Associate Professor of Genetics and Molecular Biology in the School of Dentistry and Ribeirão Preto Medical School (USP, Campus of Ribeirão Preto, Brazil), where he is also the coordinator of the Molecular Immunogenetics Group.

Part I
Basic Principles of the Transcriptome and Its Analysis

Chapter 1
What Is the Transcriptome and How It Is Evaluated

Amanda F. Assis, Ernna H. Oliveira, Paula B. Donate, Silvana Giuliatti, Catherine Nguyen, and Geraldo A. Passos

1.1 What Is the Transcriptome, How It Is Evaluated, and What Types of RNA Molecules Exist

Strictly speaking, the *transcriptome* can be conceptualized as the total set of RNA species, including coding and noncoding RNAs (ncRNAs), that are transcribed in a given cell type, tissue, or organ at any given time under normal physiological or pathological conditions. This term was coined by Charles Auffray in 1996 to refer to the entire set of transcripts. Soon after, this concept was applied to the study of large-scale gene expression in the yeast *S. cerevisiae* (Velculescu et al. 1997; Dujon 1998; Pietu et al. 1999).

A. F. Assis · E. H. Oliveira
Molecular Immunogenetics Group, Department of Genetics, School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil

P. B. Donate
Inflammation and Pain Group, Department of Pharmacology, School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil

S. Giuliatti
Bioinformatics Group, Department of Genetics, School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil

C. Nguyen
Laboratory TAGC INSERM U1090, National Institute for Health and Medical Research, Scientific Park of Luminy, Marseille, France

G. A. Passos (✉)
Molecular Immunogenetics Group, Department of Genetics, School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil

Laboratory of Genetics and Molecular Biology, Department of Basic and Oral Biology, School of Dentistry of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil
e-mail: passos@usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_1

However, due to the importance of messenger RNAs (mRNAs), which represent protein-coding RNAs, the term transcriptome is often associated with this set of RNA and as an analogy species. Researchers later coined the analogous term *miR-Nome* to refer to the total set of miRNAs.

The *proteome* is conceptually similar to the transcriptome and refers the total set of proteins translated in a given cell type, tissue, or organ at any given time during normal physiological or pathological conditions. Due to its importance, and as a consequence of the transcriptome, the proteome will be discussed on Chap. 8 of this book. Moreover, we suggest the following reviews for further reading: Alfaro et al. 2021; Foreman et al. 2021; Joyce and Ternette 2021; Anderson 2014; Forler et al. 2014; Padron and Dormont 2014; Altelaar et al. 2013; and Ahrens et al. 2010.

Analyses of the transcriptome began well before its conceptualization. Large-scale analyses of gene expression in the murine thymus gland (Nguyen et al. 1995), the human brain and liver (Zhao et al. 1995), and human T cells (Schena et al. 1996) have been performed since the mid-1990s. These independent groups used cDNA clones arrayed on nylon membranes or glass slides to hybridize labeled tissue- or cell-derived samples. These arrayed cDNA clones represented the prototypes of the modern microarrays currently used in transcriptome research (Jordan 2012).

1.1.1 How the Transcriptome Is Evaluated: The Birth of Transcriptome Methods

Although the first method used to analyze transcriptional gene expression emerged in 1980 with the development of Northern blot hybridization (Wreschner and Hersberg 1984), this method was not and still is not capable of being performed on a large scale, and thus cannot be considered a transcriptome approach. In 1990s, the human genome project, through partially automated DNA sequencing, had the ambition to identify, characterize, and analyze all of the genes in the human genome (Watson 1990; Cantor 1990). This revolutionary approach led to thousands of entries that were constructed via the tag-sequencing of randomly selected cDNA clones (Adams et al. 1991, 1992, 1993a, b; Okubo et al. 1992; Takeda et al. 1993), thus opening an avenue for high-throughput approaches by making these data widely available in repositories such as the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST). As more and more genes are identified, efforts are now being redirected toward understanding the precise temporal and cellular control of gene expression. The advances provided by the current progress in high-throughput technologies have enabled the simultaneous analysis of the activity of many genes in cells and tissues, essentially depicting a molecular portrait of the tested sample. The transcriptome approach, based on the large-scale measurement of mRNA, became the method of choice among the emerging technologies of the so-called functional genomics, primarily because this method was rapidly identified as one that can be performed at a reasonably large scale using highly parallel hybridization methods, and it has allowed a more holistic view of what is really happening in the cell (Sudo et al. 1994; Granjeaud et al. 1996, 1999; Botwell 1999; Jordan 1998).

As mentioned above, the first transcriptome analysis was performed on large nylon arrays using high-density filters containing colony cDNA (or PCR products) followed by quantitative measurements of the amount of hybridized probe at each spot. A common platform used spotted cDNA arrays, where cDNA clones representing genes were robotically spotted on the support surface either as bacterial colonies or as PCR products. These "macroarrays," or high-density filters, were made on nylon membranes measuring approximately 10 cm^2. Although this is now considered a dated approach, it was nonetheless effective enough to test sets of hundreds or even a few thousand genes.

DNA arrays allow the quantitative and simultaneous measurement of the mRNA expression levels of thousands of genes in a tissue or cell sample. The technology is based on the hybridization of a complex and heterogeneous RNA population derived from tissues or cells. Initially, this was referred as a "complex probe," i.e., a complex mix that contains varying amounts of many different cDNA sequences, corresponding to the number of copies of the original mRNA species extracted from the sample. This complex probe was produced via the simultaneous reverse transcription and ^{33}P labeling of mRNAs, which were then hybridized to large sets of DNA fragments, representing the target genes, arrayed on a solid support. Thus, each individual experiment provided a very large amount of information (Gress et al. 1992; Nguyen et al. 1995; Jordan 1998; Velculescu et al. 1997; Zhao et al. 1995; Bernard et al. 1996; Pietu et al. 1996; Rocha et al. 1997).

1.1.2 *Miniaturization: An Obvious Technological Evolution Toward Microarrays*

One of the major challenges that researchers faced was to obtain the highest possible sensitivity when working with a limited amount of sample (biopsies, sorted cells, etc.). In this regard, five parameters were taken into account: (1) the amount of DNA fixed on the array support; (2) the concentration of RNA that should be labeled with the ^{33}P isotope; (3) the specific activity of the labeling; (4) the duration of the hybridization; and (5) the duration of exposure of the array to the phosphor imager shields.

The miniaturization of this method lay in the intrinsic physical characteristics of nylon membranes, which allowed a significant increase in the amount of immobilized DNA. The feasibility of miniaturizing nylon was demonstrated in the Konan Peck (Academia Sinica, Taiwan) laboratory in 1998 using a colorimetric method as the detection system (Chen et al. 1998). A combination of nylon microarrays and ^{33}P-labeled radioactive probes was subsequently shown to provide similar levels of sensitivity compared with the other systems available at the time, making it possible to perform expression profiling experiments using submicrogram amounts of unamplified total RNA extracted from small biological samples (Bertucci et al. 1999).

These observations had important implications for basic and clinical research in that they provided a cheaper alternative approach that was particularly suitable for groups operating in academic environments and led to a large numbers of expression profiling analyses when only small amounts of biological material were available.

Microarrays based on solid supports, typically coated glass, were simultaneously developed in different academic and industrial laboratories. These arrays boasted the advantage of performing dual hybridization of a test sample and a reference sample, as they could be labeled with two different fluorescent compounds, namely the fluorochrome "Cy-dyes" cyanine-3 (Cy3) and cyanine-5 (Cy5) (Chee et al. 1996).

Around the same time, another well-known DNA array platform was developed by Affymetrix (Santa Clara, CA, USA). Their array used oligonucleotide chips featuring hundreds of thousands of oligonucleotides that were directly synthesized in situ on silicon chips (each measuring a few cm^2) using photochemical reactions and a masking technology (Lockhart et al. 1996). This microarray platform promised a rapid evolution in miniaturization because it was based on the synthesis of short nucleic acid sequences, which could be updated on the basis of the current knowledge of the genome.

It quickly became clear in the academic community, as well as in industry, that the available microarray technologies represented the beginning of a revolution with considerable potential for applications in the various fields of biology and health because gene function is one of the key elements that researchers want to extract from a DNA sequence. Microarrays have become a very useful tool for this type of research (Gershon 2002). Therefore, the development of the microarray opened the door to various DNA chip technologies based on the same basic concept. For example, the maskless photolithography used to produce oligonucleotide arrays was originally developed in 1999 using the light-directed synthesis of high-resolution oligonucleotide microarrays with a digital micromirror array to form virtual masks (Singh-Gasson et al. 1999). However, this technology was barely accessible to academic laboratories at the time because of the high initial cost, the limited availability of equipment, non-reusability, and the need for a large amount of starting RNA (Bertucci et al. 1999).

This development formed the basis for the NimbleGen company, which in 2002 demonstrated the chemical synthesis quality of maskless arrays synthesis (MAS) and its utility in constructing arrays for gene expression analysis (Nuwaysir et al. 2002). Currently, Roche-NimbleGen is focused on products for sequencing (https://sequencing.roche.com/en.html/).

Similarly, in 2005, Edwin Southern's team developed a method for the in situ synthesis of oligonucleotide probes on polydimethylsiloxane (PDMS) microchannels through the use of conventional phosphoramidite chemistry (Moorcroft et al. 2005). This became the basis of the Oxford Gene Technology company (http://www.ogt.co.uk/), which today develops array products centered on cytogenetics, molecular disorders, and cancer.

It is also widely known that Affymetrix (https://www.thermofisher.com/br/en/home/life-science/microarray-analysis.html/) and Agilent (http://www.home.

agilent.com/agilent) developed the most popular microarray technology for expression profiling based on ink jet technology, which is still widely available in the transcriptome market.

1.1.3 Reliable Microarray Results Depend on a Series of Complex Steps

The reliability of transcriptome results has concerned scientists since the beginning of transcriptome research, resulting in a number of studies comparing the different platforms, which was a real challenge in the early 2000s. Transcriptomic results largely depend on the technology used, which itself is dependent on several complex steps, ranging from the fabrication of the microarray to the experimental conditions, in addition to the chosen detection system, which also determines the method of analysis.

The results obtained with one microarray platform cannot necessarily be reproduced on another, and differences in the presence of different target sequences representing the same gene on different arrays can make it extremely difficult to integrate, combine, and analyze the data (Järvinen et al. 2004).

The fabrication of high-quality microarrays has been a challenging task, taking a decade to reach several stabilized solutions, and has become an industry of its own. There are a large number of parameters and factors that affect the fabrication of a microarray, as performance depends on the array geometry, chemistry, and spot density, as well as on characteristics such as morphology, probe and hybridized density, background, and sensitivity (Dufva 2005). Among the different methods used to fabricate DNA microarrays, in situ synthesis is the most powerful because a very high spot density can be achieved and because the probe sequence can be chosen for each synthesis.

To achieve a 10^5-fold dynamic range, which is an important parameter for gene expression analysis, the spots must contain at least 10^5 molecules, and the optimal spot size should be large enough to acquire the maximum hybridized density to obtain good sensitivity. Bead arrays that have different combinations of fluorescent dyes, which essentially constitute a barcode tag associated with the different immobilized probes, appeared to be the next evolution because they are in suspension and are therefore suitable for automation using standard equipment, leading to extremely high-throughput approaches. Optical microarrays that are detected via flow cytometry can use a large number of different beads because each bead can be decoded using a series of hybridization reactions following the immobilization of the beads to the optical fibers (Ferguson et al. 2000; Epstein et al. 2003). This increases the multiplex capacity to several thousands of different beads (Gunderson et al. 2004). Optical fiber microarrays have been commercialized by Illumina (http://www.illumina.com/), currently the leader in high-throughput sequencing technology, which allow the measurement of expression profiles by counting the amount of each RNA molecule expressed in a cell.

Experimental conditions also vary from lab to lab, as the preparation is dependent on the array platform. Variations in the quality of RNA preparations can be evaluated using the 2100 Bioanalyzer instrument developed by Agilent, which has become a standard, even if some slight variations have been observed from time to time. This system provides sizing, quantitation, and quality control for RNA and DNA, as well as for proteins and cells, on a single platform, providing high-quality digital data (https://www.agilent.com/en/product/automated-electrophoresis/bioanalyzer-systems/bioanalyzer-instrument/2100-bioanalyzer-instrument-228250) (Fig. 1.1).

The preparation of RNA prior to hybridization can affect microarray performance, particularly in terms of data accuracy, by distorting the quantitative measurement of transcript abundance. To obtain enough material from an initial nano- or picogram range of starting material, the RNA is transcribed in vitro and amplified using different protocols, which can introduce bias. In 2001, several publications discussed the different commercial protocols that were available. A publication from Charles Decreane's team examined the methods for amplifying picogram amounts of total RNA for whole genome profiling. The authors set up a specific experiment to compare three commercial RNA amplification protocols, Ambion messageAmp™, Arcturus RiboAmp™, and Epicentre Target Amp™, to the standard target labeling procedure proposed by Affymetrix, and all of the samples were tested on Affymetrix GeneChip microarrays (Clément-Ziza et al. 2009). The results obtained in this study indicated large variations between the different protocols, suggesting that the same amplification protocol should always be used to maximize the comparability of the results. Additionally, it was found that the RNA amplification affects the expression measurements as well, which was in agreement with earlier observations seen at the nanogram scale, as well as with other studies that were concerned with this question (Nygaard and Hovig 2006; Singh et al. 2005; Wang et al. 2003; Van Haaften et al. 2006; Degrelle et al. 2008).

In 2012, questions surrounding RNA amplification were still relevant. Indeed, even if the amplification of a small amount of RNA is reported to have a high reproducibility, there is still bias, and this can become time consuming. Even taking into account a correlation coefficient of 0.9 between microarray assays using non-amplified and qRT-PCR samples, the matter should still be reconsidered. In one study, the authors used the 3D-Gene™ microarray platform and compared samples prepared using either a conventional amplification method or a non-amplification protocol and a probe set selected from the MicroArray Quality Control (MAQC) project (https://www.fda.gov/science-research/bioinformatics-tools/microarraysequencing-quality-control-maqcseqc/). They found that the samples from the non-amplification procedure had a higher quantitative accuracy than those from the amplification method but that the two methods exhibited comparable detection power and reproducibility (Sudo et al. 2012).

However, in the above study, the researchers also used a few micrograms of RNA and a large volume of hybridization buffer. It is known that the ability to reduce the quantity of input RNA while maintaining the reaction concentration can be achieved in a device that decreases the hybridization reaction volume. Devices developed for use with beads have this characteristic; therefore, would hybridization using a bead device resolve this issue?

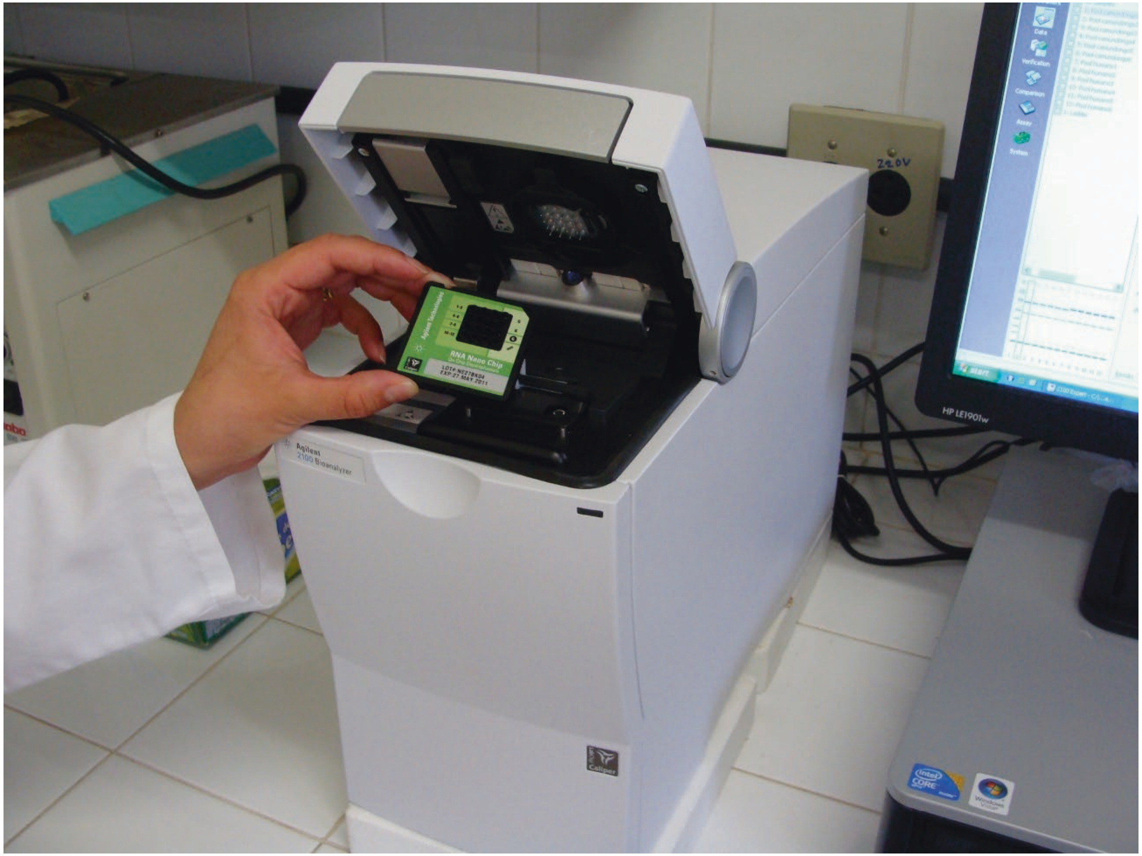
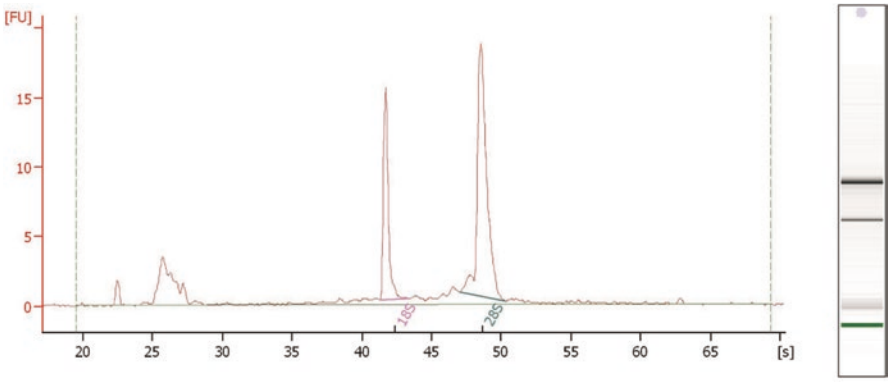


Fig. 1.1 Agilent Bioanalyzer model 2100 showing in (**a**) a RNA Nano Chip and in (**b**) a typical result of a microfluidic electrophoresis of a total human RNA sample extracted from leukocytes. On the right side of this figure appears a virtual gel with the respective bands of 28S and 18S rRNAs and 5S rRNA plus 4S tRNAs (from top to bottom). On the left side is shown the densitometry of this gel where appear the respective peaks of 28S rRNA, 18S rRNAs, 5S rRNA, and 4S tRNAs. The rRNA ratio (28S/18S) = 2.0 enabled a RNA integrity number (RIN = 9.7), which indicated that this sample was intact (not degraded)



Overall Results for sample 1

RNA Area:	82.3	RNA Integrity Number (RIN):	9.7 (B.02.07)
RNA Concentration:	118 ng/µl	Result Flagging Color:	🟪
rRNA Ratio [28s / 18s]:	2.0	Result Flagging Label:	RIN: 9.70

Fig. 1.1 (continued)

1.1.4 Bioinformatics and Standardization Approaches: A Possible Solution?

With regard to bioinformatics and standardization approaches, the MAQC project was initiated in 2006 to address these questions, as well as other performance and data analysis issues. The Microarray Quality Control (MAQC Consortium 2006) (https://www.fda.gov/science-research/bioinformatics-tools/microarraysequencing-quality-control-maqcseqc/) study tested a large number of laboratories, platforms, and samples and found that there were notable differences in various dimensions of performance between microarray platforms. Each microarray platform has different trade-offs with respect to consistency, sensitivity, specificity, and ratio compression. One interesting result was that platforms with divergent approaches for measuring expression often generated comparable results. The authors of this study concluded that the technical performance of microarrays supports their continued use for gene expression profiling in basic and applied research and may lead to the use of microarrays as a clinical diagnostic tool as well. This project has provided the microarray community with standards for data reporting, common analysis tools, and useful controls that can help promote confidence in the consistency and reliability of these gene expression platforms (MAQC Consortium 2006). Similarly, in 2007, another meta-analysis of microarray results suggested several recommendations for standardization under the Standard Microarray Results Template (SMART) to facilitate the integration of microarray studies and proposed the implementation of the Minimum Information About a Microarray Experiment (MIAME) currently at Functional Genomics Data Society (http://fged.org/) to facilitate the comparison of results (Cahan et al. 2007).

Given that measurement precision is critical in clinical applications, the question of the measurement precision in microarray experiments was addressed again in 2009 through an inter-laboratory protocol. In this study, the authors analyzed the results of three 2004 Expression Analysis Pilot Proficiency Test Collaborative studies using different methods. The study involved 13 participants out of 16, each of whom provided triplicate microarray measurements for each of two reference RNA pools. To facilitate communication between the user and developer, this study sought to set up standardized conceptual tools, but the result of this analysis was relatively disappointing and did not allow the creation of a gold standard, though it did put forth several recommendations (Duewer et al. 2009).

All of these studies focus on the same concept that has been defended since 2001 by the Microarray Gene Expression Data Society, now Functional Genomics Data Society (http://fged.org/) – the reanalysis and reproduction of results by the scientific community. The MGED society was the first to define the MIAME, which describes the minimum information required to ensure that microarray data can be easily interpreted and that the results derived from their analysis can be independently verified. This protocol became the standard for recording and reporting microarray-based gene expression data and for inserting it in databases and public repositories (Brazma et al. 2001; Ball et al. 2002). Currently, raw and/or normalized microarray data are deposited either in the ArrayExpress databank (https://www.ebi.ac.uk/arrayexpress/) or in the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/), providing the scientific community with data for further analysis.

1.1.5 Analysis of the Expression Data

The past two decades have seen the development of methods that allow for a nearly complete analysis of the transcriptome, in the form of microarrays and, more recently, RNA-Seq, which are the most popular technologies used in genome-scale transcriptional studies. These high-throughput gene expression analysis systems generate large and complex datasets, and the development of computational methods to obtain biological information from the generated data has been the primary challenge in bioinformatics analysis.

Next Generation Sequencing (NGS) technology has experienced a great technical advance and a decrease in costs lately. In this way, it is undeniable that RNA-Seq has become the most used tool for comprehensive identification and characterization of both coding and noncoding RNAs from bulk tissue and/or cells and even specific cell types through single-cell sequencing (Stark et al. 2019).

Minnier et al. (2018) point that a gene expression profile experiment is composed of five related components: (1) study design, (2) sample collection and processing, (3) data generation, (4) data analysis, and (5) data interpretation. In selecting a particular technology or platform for the investigation of the transcriptome, the biological question must be considered, as well as the characteristics of the sample under study (fresh or preserved material; type, yield and quality of the extracted RNA) and the capabilities of the available analysis platforms. If researchers are concerned in measure differential expression or biological pathways that are changing under the study conditions, any platform which provides a comprehensive measurement of mRNAs should be adequate. If a broader measurement of RNA transcripts, like miRNAs or lncRNAs, is desired, microarrays and RNA-Seq technologies do, and this last one provides opportunity for discovery of unannotated transcripts. Microarray relies on the availability of sequence information and gene annotation for design and synthesis of probes, an issue that is not a point for human studies and widely used model organisms and agriculturally important species.

Concerning analysis, microarrays and RNA-Seq are genome-wide gene expression profiling technologies, so they both generate a large amount of data, which places certain demands on the analysis software. Fortunately, microarrays have benefited from the availability of many commercial and open-source software packages for data manipulation that have been developed over the years. RNA-Seq, however, demands more bioinformatics expertise. There are publicly available online tools such as the Galaxy platform (Goecks et al. 2010), but a basic knowledge of UNIX shell programming and Perl/Python scripting is necessary for data modification. Furthermore, similar to microarray analysis, a familiarity with the R programming environment is useful, as the software programs for many of the downstream analyses are collected in the Bioconductor (http://www.bioconductor.org/) (Gentleman et al. 2004) suite of the R package. Other important considerations regarding the choice for RNA-Seq include the need for data storage resources and computing systems with large memories and/or many cores to run parallel, sophisticated algorithms efficiently and faster.

In this section, we present the main steps for analyzing multi-dimensional genomic data derived from the application of microarray or RNA-Seq assays based on a common pipeline illustrated in Fig. 1.2.

1.1.5.1 Experimental Design

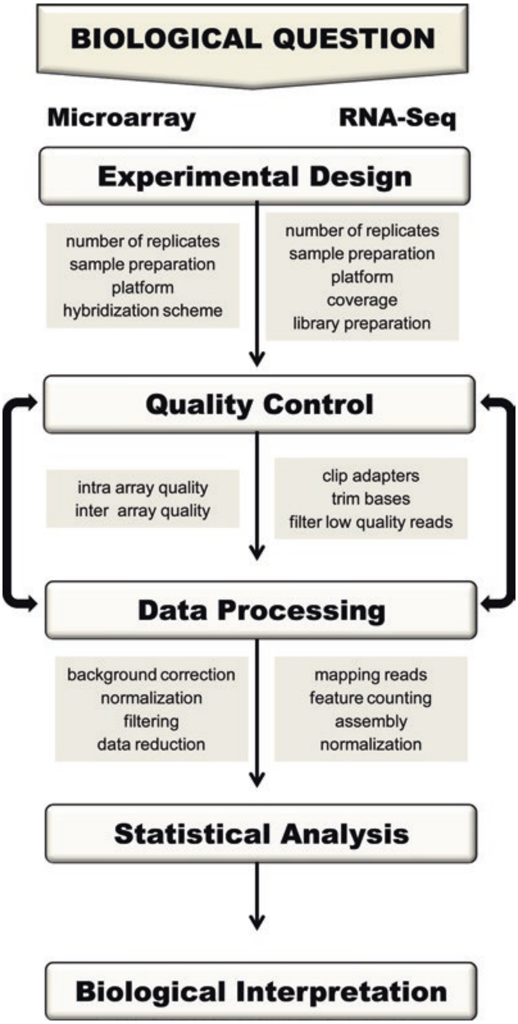
The aim of the experimental design is to make the experiment maximally informative given a certain number of samples and resources and to ensure that the questions of interest can be answered. All of the decisions made at this initial step will affect the results of all the subsequent steps. The consequences of an incorrect or poor design range from a loss of statistical power and an increased number of false negatives to the inability to answer the primary scientific question (Stekel 2003).

The basic principles of experimental design rely on three fundamental aspects formalized by Fisher (1935), namely, replication, randomization, and blocking.

Randomization dictates that the experimental subjects should be randomly assigned to the treatments or conditions to be studied to eliminate unknown factors that may potentially affect the results (Fang and Cui 2010).

A relevant issue in RNA-Seq experiments is sequencing depth or library size, which is the number of sequenced reads for a given sample. An optimal sequencing

Fig. 1.2 An overview of the steps in a typical gene expression microarray or RNA-Seq experiment



depth depends on the goals of the experiment and optimal library size depends on the complexity of the targeted transcriptome (Tarazona et al. 2011; Conesa et al. 2016; Stark et al. 2019). By evaluating different algorithms available for the analysis of RNA-Seq, Tarazone and colleagues (2011) verified that methodologies suffer from a strong dependency on sequencing depth for their differential expression calls since deep sequencing improves quantification and identification but might result in a considerable number of false positives (transcriptional noise and off-target transcripts) that increases as the number of reads grows.

Replication is essential for estimating and decreasing the experimental error and, thus, to detect the biological effect more precisely. A true replicate is an independent repetition of the same experimental process and an independent acquisition of the observations. There are different levels of replication in gene expression experiments: (1) a technical replicate provides measurement-level error estimates and (2) a biological replicate provides estimates of the population-level variability. If the goal is to evaluate the technology, technical replicates alone are sufficient. Otherwise, if the goal is to investigate the biological differences between tissues/conditions/treatments, biological replicates are essential (Alison et al. 2006; Fang and Cui 2010). Replication is widely used in microarray experiments, though technical replicates are generally no longer performed, as analyses have shown that the results will be relatively consistent overall (Slonin and Yanai 2009). In RNA-Seq studies, the number of replicates that should be included is determined by three factors: (i) the amount of technical variability in the RNA-Seq procedures, (ii) biological variability of the system under study, and (iii) the desired statistical power. The first one is influenced by the technical noise, mainly RNA extraction and library preparation, and the biological variation. Biological variation is particular to each experimental system; nevertheless, a minimum of three replicates is necessary for any inference on the population analysis. Regarding the third factor, statistical power, a proper analysis requires estimates of the within-group variance and gene expression levels; these can be done by software packages that provide a theoretical estimate of power over a range of variables considering the method used for differential expression analysis (Conesa et al. 2016). Setting a number of appropriate replicas for a RNA-Seq experiment is not always simple, Lamarre et al. (2018) suggested four replicates for DEGs analysis, but they point the necessity of measuring biological before settling on an appropriate number of replicates. Stark et al. (2019) also emphasize that for highly diverse samples many more replicates are likely to be required in order to pinpoint changes with confidence.

As with microarray studies, RNA-Seq experiments can be affected by the variability coming from nuisance factors, often called technical effects, such as the processing date, technician, reagent batch, and the hybridization/library preparation effect (Fig. 1.3). In addition to these effects, in RNA-Seq experiments, there are also other technology-specific effects, for example, there is variation from one flow cell to another and variation between the individual lanes within a flow cell due to systematic variation in the sequencing cycling and/or base calling (Alison et al. 2006; Slonin and Yanai 2009; Auer and Doerge 2010; Fang and Cui 2010; Luo et al. 2010; Conesa et al. 2016; Chatterjee et al. 2018).

In the case of microarray and RNA-Seq experiments, design issues are intrinsically dependent on hybridization and library construction, respectively. It is beyond the scope of this section to discuss and compare the different technologies available, but we recommend reading the following articles for microarray technologies: Patersen et al. (2006), Allison et al. (2006), Stekel (2003), Churchill (2002), Kerr and Churchill (2001), and Jordan (2012). For RNA-Seq technologies, see Auer and Doerge (2010), Fang and Cui (2010), Van Dijk et al. (2014), SEQC/MAQC-III Consortium (2014), Conesa et al. (2016), Chatterjee et al. (2018), Stark et al. (2019), as well as Chaps. 3 and 5 of this book.

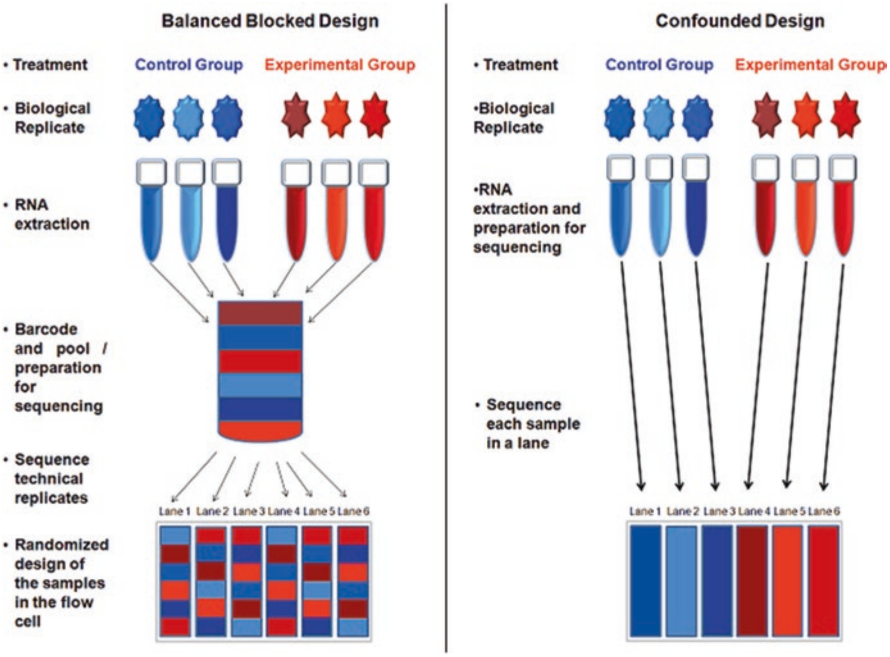

Balanced Blocked Design
• Treatment
Control Group
Experimental Group
• Biological Replicate
• RNA extraction
• Barcode and pool / preparation for sequencing
• Sequence technical replicates
• Randomized design of the samples in the flow cell
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6
Confounded Design
• Treatment
Control Group
Experimental Group
•Biological Replicate
•RNA extraction and preparation for sequencing
• Sequence each sample in a lane
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6

Fig. 1.3 Comparison of two methods for testing differential expression between treatments (**a**) (red) and (**b**) (blue). In the ideal balanced block design (left), six samples are barcoded, pooled, and processed together. The pool is then divided into six equal portions that are input into six flow cell lanes. The confounded design (right) represents a typical RNA-Seq experiment and consists of the same six samples, with no barcoding, and does not permit batch and lane effects to be distinguished from the estimate of the intra-group biological variability. (Adapted from Auer and Doerge 2010)

1.1.5.2 Quality Control

To assure the reproducibility, comparability, and biological relevance of the gene expression data generated by high-throughput technologies, several research groups have provided guidelines regarding quality control (QC) (Fig. 1.4):

- Minimum Information About a Microarray Experiment (MIAME): describes the minimum information required to ensure that microarray data can be easily interpreted and that the results derived from their analysis can be independently verified (Brazma et al. 2001).
- External RNA Control Consortium (ERCC): develops external RNA controls useful for evaluating the technical performance of gene expression assays performed by microarray and qRT-PCR (Baker et al. 2005).
- MicroArray Quality Control (MAQC) Consortium: a community-wide effort, spearheaded by the Food and Drug Administration (FDA), that seeks to experimentally address the key issues surrounding the reliability of microarray and next-generation sequencing technologies. Now in its fourth phase (MAQC-IV),

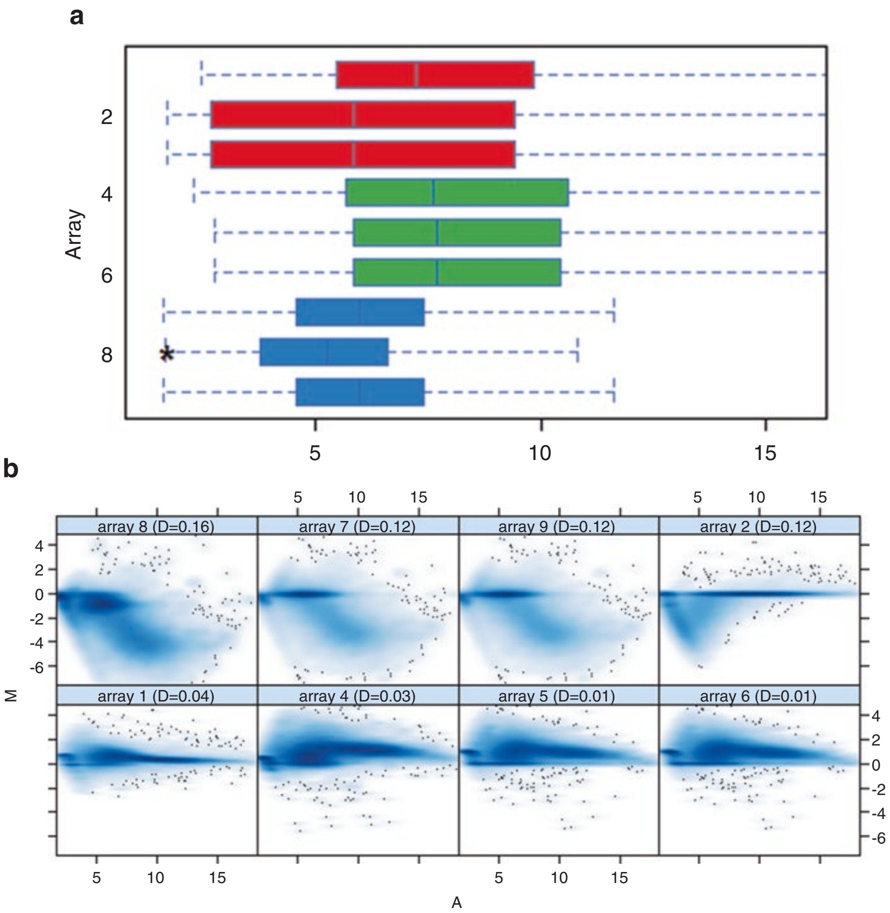


Fig. 1.4 Quality control plots of raw data sets. (**a**) Boxplots presenting various statistics for a given data set. The plots consist of boxes with a central line and two tails. The central line represents the median of the data, whereas the tails represent the upper (75th percentile) and lower (25th percentile) quartiles. These plots are often used to describe the range of log ratios that is associated with replicate spots. (**b**) *MA* plots are used to detect artifacts in the array that are intensity dependent

also known as Sequencing Quality Control Phase 2 (SEQC2), the MAQC project consists of three specific aims: (1) to develop quality metrics for reproducible NGS results from both whole genome sequencing (WGS) and targeted gene sequencing (TGS), (2) to benchmark bioinformatics methods for WGS and TGS toward the development of standard data analysis protocols, and (3) to assess the joint effects of key parameters affecting NGS results and interpretation for clinical application (Shi et al. 2006, 2010; SEQC/MAQC-III Consortium 2014) (https://www.fda.gov/science-research/bioinformatics-tools/microarraysequencing-quality-control-maqcseqc).

- Standards, Guidelines and Best Practices for RNA-Seq: a guideline for conducting and reporting on functional genomics experiments performed with RNA-Seq. It focuses on the best practices for creating reference-quality transcriptome measurements (The ENCODE Consortium 2011, 2016) (https://www.encodeproject.org/about/experiment-guidelines/).

However, there are several sources of variability originating from biological and technical causes that can affect the quality of the resulting data, including biological heterogeneity in the population, sample collection, RNA quantity and quality, technical variation during sample processing, and batch effects, among others. Some of these issues can be avoided with an appropriate and carefully designed experiment that controls for the different sources of variation, but others require a quality assessment of the raw data through computational support tools. Therefore, regardless of the technology used to measure gene expression, ensuring quality control is a critical starting point for any subsequent analysis of the data (Irizarry et al. 2005; Heber and Sick 2006; Conesa et al. 2016; Minnier et al. 2018; Chatterjee et al. 2018).

With regard to microarray technology, many tools applying diagnostic plots have been developed to visualize the spread of data and compare and contrast the probe intensity levels between the arrays of the dataset. These qualitative visualization plots include histograms, density plots, boxplots, scatter plots, MA plots, score plots of the PCA, hierarchical clustering dendrograms, and even chip pseudo plots and RNA degradation plots. Comparing the probe intensity between samples allows us to observe if one or more of the arrays have intensity levels that are drastically different from the other arrays, which may indicate a problem with the arrays. For a better review of the use of diagnostic plots in quality control metrics, please see Gentleman et al. (2005) and Heber and Sick (2006).

Concerning RNA-Seq, several sequence artifacts are quite common, including read errors (base calling errors and small indels), poor quality reads, and adaptor contamination. Such artifacts need to be removed before performing downstream analyses; otherwise they may lead to erroneous conclusions. Performing a quality assessment of the reads allows us to determine the need for filtering (or cleaning) the data, removing low quality sequences, trimming bases, removing linkers, determining overrepresented sequences, and identifying contamination or samples with a low sequence performance. The most important parameters used to verify the quality of the raw sequencing data are the base quality, the GC content distribution, and the duplication rate (Guo et al. 2013; Patel and Jain 2012).

In addition to the QC pipelines provided commercially by the sequencing platform, there are online/standalone software packages and pipelines available as well (see: http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools). These packages present different features, and many are designed for a particular sequencing platform, such as NGS QC for the Illumina and Roche 454 platforms (Patel and Jain 2012) or for a specific data storage format, such as FastQC toolkit and FastQScreen, which were both developed by the Brabaham Institute. The FastQC (Fig. B) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and FASTX-Tool kits (http://hannonlab.cshl.edu/fastx_toolkit/) include many of the tools used to remove indexes, barcodes, and adapters and filter out the reads based on the quality metrics of the FASTQ files. For a comparison of some of the available QC tools for RNA-Seq, please refer to Patel and Jain (2012).

1.1.5.3 Data Processing

Once the quality of the data has been assessed and the applicable changes have been done, it is still necessary to make some processing before the analysis of differential expressed genes. The main objective in processing raw data is to remove unwanted sources of variation, ensuring the accuracy of the final results. There are several different methods to process data being assayed and the form to perform it depends on how data were generated.

Essentially, microarray processing involves three steps depending on the type of array: (1) background adjustment, which divides the measured hybridization intensities into a background, and a signal component; (2) summarization which combines probe-level data into gene expression values, reducing multiple probes representing a single transcript to a single measurement of expression; and (3) normalization which has the aim of removing non-biological variation between arrays (Heber and Sick 2006). Other potential processing steps include transformation of data from the raw intensities into log intensities and data filtering by removing flagged features which are features for which the image-processing software has detected some type of problem (Stekel 2003; Allison et al. 2006).

Microarray data must be background corrected to remove signal due to nonspecific hybridization or spatial heterogeneity across the array. The background is a measure of the ambient signal obtained, generally, from the mean or median of the pixel intensity values surrounding each spot (Ritchie et al. 2007). The traditional correction is to subtract the local background measures from the foreground values but the main problem with this procedure is that it could give negative correct intensities and high variability of low intensity log-ratios when the background is higher than the feature intensity (Stekel 2003). Different methods have been developed as alternatives, for example the empirical Bayes model developed by Kooperberg et al. (2002), setting a small threshold value as suggested by Edwards (2003), the variance stabilization method (Vsn) (Huber et al. 2002), the normexp (normal-exponential convolution) method implemented by the RMA algorithm (Irizarry et al. 2003), and MLE (maximum likelihood estimation for normexp) (Silver et al. 2009). A comparison of several methods can be accessed on Ritchie et al. (2007).

Microarray signal intensity normalization has been widely used to adjust for experimental artifacts within array and between all the samples so that meaningful biological comparison can be made (Quackenbush 2001; Luo et al. 2010). According to Stekel (2003) the methods for may be broadly classified into two methods:

1. Within array normalization (normalize the M-values for each array separately) – methods applicable in two-channel arrays which aim is to adjust Cy3 and Cy5 intensities into equal footing. Methods as linear regression of Cy5 against Cy3,

linear regression of log ratio against average intensity and non-linear (Loess) regression of log ratio against average intensity can correcting for different response of Cy3 and Cy5 channels. However, these methods rely on the assumption that the majority of the genes on the microarray are not differentially expressed. If this assumption is not true, a different normalization method, using a reference sample for example, would be more appropriate.

2. Between array normalization (normalize intensities or log-ratios to be comparable across arrays) – used for one and two channel arrays. Various methods have been proposed in this approach, for example scaling to mean or median, centering, and quantile. Bolstad et al. (2003) presented a review of some methods and found quantile normalization to perform favorable.

After processing procedure, it is strongly recommended to check the performance of the choosing method; it can be done applying the diagnostic plots cited above at Quality Control session. Several studies have been published concerning performance of the processing methods (Bolstad et al. 2003; Ploner et al. 2005) but most studies find Robust Multichip Average (RMA) (Irizarry et al. 2003) to be among the best methods. It applies a model-based background adjustment followed by quantile normalization and a robust summary method (median polish) on the log2 intensities to obtain probe set summary values.

RNA-Seq data processing steps considered in our pipeline are: (1) alignment and assembly of sequencing reads, (2) quantification of transcript abundance, and (3) filtering lowly expressed features and normalization of read counts.

A common characteristic of all high-throughput sequencing technologies is the generation of relatively short reads which should be mapped to a reference sequence, being a reference genome or a transcriptome database. This is a critical task for most applications of the technology because the alignment algorithm must be able to efficiently find the right location of each read from a potentially large quantity of reference data (Fonseca et al. 2012). The assembly of the transcriptome consists in the reconstruction of the full-length transcripts, except in the case of small classes of RNA that are shorter than the sequencing length and no require assembly. The methods used to assembly reads fall into two main classes: (1) assembly based on a reference genome and (2) de novo assembly (Martin and Wang 2011). Alignment tools, such as TopHat (Kim et al. 2013), STAR (Dobin et al. 2013), or HISAT (Kim et al. 2015), rely on a reference genome and perform a spliced alignment allowing for gaps in the reads when compared to the reference genome. The strategies to map the reads and assemble the transcriptome and the available tools will be presented in more detail in Chap. 3.

Quantification in RNA-Seq experiments stands for assign mapped reads to genes or transcripts, to determine abundance measures and this step is the basis for to estimate gene and transcript expression. This application is primarily based on the number of reads that overlap known genes, using a transcriptome annotation (Rapaport et al. 2013; Conesa et al. 2016; Stark et al.2019). Commonly used quantification tools include RSEM (Li and Dewey 2011), CuffLinks (Trapnell et al. 2012), and HTSeq (Anders et al. 2015). Usually, the results are combined into an expression matrix, with a row for each expression feature (gene or transcript) and a column for each sample, with the values being either actual read counts or estimated abundances (Stark et al. 2019).

Normalization should be applied in read counts due to two main sources of systematic variability: (i) RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample; and (ii) the variability in the number of reads produced for each run causes fluctuation in the number of fragments mapped across samples. Properly normalization will enable accurate comparison of expression levels between and within samples (Garber et al. 2011; Dillies et al. 2013). Filtering step is used to remove features with uniformly low read abundance has been shown to improve the detection of true differential expression (Bourgon et al. 2010). The RPKM (reads per kilobase of transcript per million mapped reads) was the widely used normalization metric. It normalizes a transcript read count by both its length and the total number of mapped reads in the sample (Mortazavi et al. 2008). This approach facilitates comparisons between genes within a sample and combines between and within sample normalization. When data originate from paired-end sequencing, the FPKM (fragments per kilobase of transcript per million mapped reads) metric is used (Garber et al. 2011; Dillies et al. 2013). RSEM algorithm uses an expectation maximization approach that returns transcripts per million (TPM) values. The transcript fraction measure is preferred over the popular RPKM and FPKM measures because it is independent of the mean expressed transcript length and is thus more comparable across samples and species (Li and Dewey 2011).

Normalization methods for RNA-Seq that correct for more subtle differences between samples by applying inter-sample normalization by scaling factors, such as quartile or median, have been proposed: (i) Total count (TC): in which gene counts are divided by the total number of mapped reads (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset; (ii) Upper Quartile: which is very similar in principle to TC, the total counts are replaced by the upper quartile of counts different from 0 in the computation of the normalization factors; (iii) Median: also similar to TC, the total counts are replaced by the median counts different from 0 in the computation of the normalization factors; (iv) DESeq: normalization method included in the DESeq Bioconductor package (version 1.6.0) that is based on the hypothesis that most genes are not differentially expressed (DE); (v) Trimmed Mean of M-values (TMM): normalization method implemented in the edgeRBioconductor package (version 2.4.0). It is also based on the hypothesis that most genes are not DE; (vi) Quantile: first proposed in the context of microarray data, this normalization method consists in matching distributions of gene counts across lanes (Dillies et al. 2013; Conesa et al. 2016; Stark et al. 2019).

1.1.5.4 Statistical Analysis and Interpretation

The main goal of gene expression studies is to determine which transcript features have changed their level of expression under some biological conditions, i.e., identify genes that are Differentially Expressed (DE) between RNA samples. DE can give insights into biological mechanisms or pathways, and form the basis for further experiments as sample and gene similarity performed in clustering analysis or testing gene set enrichment.

Differential expression analysis search for genes that have changed significantly in abundance across experimental conditions. In general, this means taking the quantified and normalized expression values for each library and performing statistical testing between samples of interest. In thesis transcript abundance of mRNA would be directly proportional to the number of reads thereby determining the expression level (Oshlack et al. 2010).

Many methods have been developed for the analysis of differential expression using microarray data. In the early days of microarray, only the simple method of fold change was used (Chen et al. 1997). But the evolution of the technique led to the necessity more accurate analysis methods and many more sophisticated statistical methods have been proposed.

Besides the traditional t-test and ANOVA approaches used to access differential gene expression in microarrays assays, methodologies with variations of these tests were created for the purpose of overcoming the problem of small sample size accessing such a large dataset: dealing with many genes but few replicates may lead to large fold changes driven by outliers, and to small error variances (Lönnstedt and Speed 2002). SAM (Significant Analysis of Microarrays) (Tusher et al. 2001) is a very popular DE method that uses a modified t-statistic to identify significant genes, using non-parametric statistics.

Other statistical approaches for microarray data analysis have introduced linear models. The bioconductor package limma, developed by Smyth (2005), applies a gene-wise linear model and allows for the analysis of complex experiments (comparing many RNA samples), as well as more simple replicated experiments with two RNA samples. Empirical Bayes and other shrinkage methods are used to borrow information across genes making the analyses stable even for experiments with small number of arrays. Another powerful method to detect DE genes in microarray experiments is rank products which is based on calculating rank products (RP) from replicate experiments and at the same time, it provides a straightforward and statistically stringent way to determine the significance level for each gene and allows for the flexible control of the false-detection rate and familywise error rate in the multiple testing situation of a microarray experiment (Breitling and Herzyk 2005).

In DE analysis methods using probability distributions have been proposed to model the count data from RNA-Seq studies: Poisson and negative binomial (NB). Poisson distribution is the basis for modeling RNA-Seq count. However, when there are biological replicates, RNA-Seq data may exhibit more variability than expected by Poisson distribution because it assumes that the variance is equal to the mean. Then it predicts smaller variation than what is seen in the data and it will be prone to high false positive rates resulting from underestimation of sampling error (Anders and Huber 2010). Assuming a negative binomial (NB) model instead of Poisson is a way to deal with this so-called overdispersed problem because the NB distribution specifies that the variance is greater than the mean (Oshlack et al. 2010; Anders and Huber 2010; Garber et al. 2011).

Statistical analysis on RNA-Seq data will be discussed more in Chap. 3. Some reviews discuss and compare statistical methods to compute differential expression; for this purpose refer to Seyednasrollah et al. (2015), Soneson and Delorenzi (2013), Rapaport et al. (2013), Conesa et al. (2016), Chatterjee et al. (2018), and Stark et al. (2019).

1.1.5.5 Classification and Enrichment Analysis

Classification can be performed a priori or a posteriori DE analysis. This process implies in either placing objects (in this case samples, genes or both) into pre-existing categories (called supervised classification) or developing a set of categories into which objects can subsequently be placed (unsupervised classification) (Allison et al. 2006). Class discovery or clustering analysis is an unsupervised classification widely used in the study of transcriptomic data because it allows us to identify co-regulates genes and/or samples with similar patterns of expression (biological classes). Various clustering techniques have been applied to the identification of patterns in gene-expression data. Most cluster analysis techniques are hierarchical; the resultant classification has an increasing number of nested classes and the result resembles a phylogenetic classification. Non-hierarchical clustering techniques also exist, such as k-means clustering, which simply partition objects into different clusters without trying to specify the relationship between individual elements (Quackenbush 2001). Einsen et al. (1998) is a classical reference on using hierarchical clustering with microarray data. The authors developed an open-source integrated pair of programs, Cluster and TreeView, for analyzing and visualizing clusters and heat maps.

Biological insight into an experimental system can be gained by looking at the expression changes of sets of genes. Many tools focusing on gene set testing, network inference, and knowledge databases have been designed for analyzing lists of DE genes from microarray datasets for example Gene Set Enrichment Analysis (Subramanian et al. 2005), DAVID (Dennis et al. 2003; Huang at al. 2009) which use of functional themes, e.g. those defined by the Gene Ontology consortium (Ashburner et al. 2000), and metabolic and signaling pathways, e.g. KEGG, (https://www.genome.jp/kegg/) (Kanehisa and Goto 2000) and Biocarta (https://maayanlab.cloud/Harmonizome/dataset/Biocarta+Pathways) combined with statistical enrichment analysis to determine, for each theme or pathway, whether it is overrepresented in a given list of DE genes. These approaches can also be applied to RNA-Seq but biases present by this kind of data, such as gene length, should be taken into account (Oshlack et al. 2010; Conesa et al. 2016). Therefore, specialized approaches and tools for enrichment analysis on RNA-Seq are being developed, for example, GO-seq (Young et al. 2010), GSASEQ (Wang and Cairns 2013), GAGE (generally applicable gene set enrichment for pathway analysis (Luo et al. 2009), Gene Set Variation Analysis (GSVA) (Hänzelmann et al. 2013), and SeqGSEA (Wang et al. 2003) packages that combine splicing and implement enrichment analyses similar to Gene Set Enrichment Analysis (GSEA).

When it comes to the transcriptome, we must not forget that a large portion of it is composed of non-protein coding transcripts. It is an emerging issue and the functional annotation of these RNAs, mainly long noncoding RNAs (lncRNAS), is challenging, even when dealing with model organisms. Researchers in this area can make use of the resources and tools present at RNAcentral (https://rnacentral.org/) that is a free, public resource that offers integrated access to a comprehensive and up-to-date set of noncoding RNA sequences provided by a collaborating group of Expert Databases representing a broad range of organisms and RNA types. Currently, in its 18th version, the RNAcentral Consortium is formed by 54 Expert Databases, and among these miRBase (https://www.mirbase.org/) contains high-quality miRNA annotations and is responsible for assigning official miRNA gene names; NONCODE (http://www.noncode.org/), an integrated knowledge database dedicated to noncoding RNAs; lncRNAdb (https://bio.tools/lncrnadb), a database providing comprehensive annotations of eukaryotic lncRNAs. The development of RNAcentral is coordinated by European Bioinformatics Institute (EMBL-EBI) (https://www.ebi.ac.uk/).

1.2 The Diversity of the Transcriptome

Unlike the genome, which is essentially static in terms of its composition and size (barring the rare occurrence of somatic and germline mutations or the rearrangement of immunoglobulin and T cell receptor genes), the transcriptome (and similarly, the miRNome) is extremely variable and depends on the phase of the cell cycle, the organ, exposure to drugs or physical agents, aging, diseases such as cancer and autoimmune diseases, and a multitude of other variables, which must be considered at the time that the transcriptome is determined. This variability arises from the fact that RNAs are differentially transcribed (or transcribed at different rates) depending on the cell type and status, though this excludes ribosomal RNAs, as they are considered housekeeping molecules.

For many years, the central dogma of molecular biology stated that RNAs molecules were intermediates between DNA and protein. This idea presupposed that the function of RNA was primarily linked to the translation of the genetic material into polypeptide chains (proteins). The genetic material was interpreted as being involved in the synthesis of these RNAs, which were termed mRNAs (Brenner et al. 1961; Jacob and Monod 1961).

During the human genome sequencing era of the 1980s and 1990s, independently led by Francis Collins and Craig Venter, the latter individual and his coworkers conceived of expressed sequence tags (ESTs), which focus on mRNAs because they encode proteins. Libraries of mRNA-derived cDNA clones were generated based on first-strand synthesis using oligonucleotide primers for that are anchored at the 3′ end of the transcript [the poly(A) tail of mRNA] (Strausberg and Riggins 2001) and then sequenced to create unique identifiers for each cDNA, with lengths ranging from 300 to 700 bp (Adams et al. 1992; Adams 2008).

ESTs were very useful for identifying new expressed genes in normal and diseased tissues (Strausberg and Riggins 2001), and transcriptome analysis at this time was largely, if not solely, based on this approach. The EST clones were distributed through the former IMAGE Consortium, whose sequences can now be retrieved via the National Center for Biotechnology Information (NCBI) dbEST Database (http://www.ncbi.nlm.nih.gov/dbEST/). The current number of public entries for all uni- or multicellular eukaryotic organisms that have been sequenced stands at more than 74 million ESTs, including more than eight million human and nearly five million mouse ESTs.

However, as was to be expected, imaginative new strategies were emerging around the same time as well. The Serial Analysis of Gene Expression (SAGE) method (Velculescu et al. 1997), which produces short sequence tags (usually 14 nucleotides in length) positioned contiguous to defined restriction sites near the 3′ end of the cDNA strand (Strausberg and Riggins 2001), has also been widely used. At the time, the NCBI created the SAGEmap as a public repository for SAGE sequences. Currently, all of the SAGE libraries have been uploaded and accessioned through the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) repository.

Another novel strategy, which had yet to be tested at that time, was the generation of open reading frame (ORF) ESTs (ORESTES). This approach was jointly developed by researchers funded by the São Paulo Research Foundation (FAPESP) and by the Ludwig Institute for Cancer Research (FAPESP/LICR)-Human Cancer Genome Project (Camargo et al. 2001). Unlike ESTs, ORESTES sequences are spaced throughout the mRNA transcript, providing a scaffold to complete the full-length transcript sequences. The authors generated a substantial volume of tags (700,000 ORESTES), which at the time represented nearly 20% of all human dbESTs (Strausberg and Riggins 2001).

The Transcript Finishing Initiative, another FAPESP/LICR project, was then undertaken for the purpose of identifying and characterizing novel human transcripts (Sogayar et al. 2004). This strategy was also novel and was based on selected EST clusters that were used for experimental validation. In this method, RT-PCR was used to fill in the gaps between paired EST clusters that were then mapped on the genome. The authors generated nearly 60,000 bp of transcribed sequences, organized into 432 exons, and ultimately defined the structure of 211 human mRNA transcripts.

However, the increasing use of modern transcriptome-wide profiling approaches, such as microarrays and whole-genome and transcriptome sequencing, allied to the precise isolation and characterization of different RNA species from eukaryotic (including mammalian) cells, led to an explosion of findings and revealed that although approximately 90% of the mammalian genome is actively transcribed into RNA molecules, only a tiny fraction (~2% of the total human genome) encodes mRNAs and, consequently, proteins (Maeda et al. 2006; Djebali et al. 2012).

In fact, the function of the genome can be seen from two different but complementary views. From a functional standpoint, only a fraction of the genome encodes RNA molecules (including coding and noncoding RNAs), and only a fraction of these are translated into proteins. In other words, when considering the genome in numerical terms, or rather the physical portion of DNA that is functional, we realize that only a small number of genes are transcribed specifically into mRNA molecules. However, a larger number of "variable" mRNA molecules are generated through alternative splicing, and these are translated into a greater number of proteins (including various isoforms). A large portion of the genome is then transcribed into noncoding RNAs, which play a role in the posttranscriptional control of mRNAs during their translation into proteins (Fig. 1.5).

Molecular mapping of the human genome has been largely resolved, revealing slightly more than three billion bp encompassing approximately 20–25,000 functional nuclear genes and mitochondrial DNA located in the cytoplasm. We suggest consulting the ENCODE Project (http://www.genome.gov/encode/) to follow ongoing progress in the identification of the functional elements in the human genome sequence. Nevertheless, the definition of the human transcriptome is still far from set, and it appears that most of the RNA molecules in eukaryotic cells are composed of ncRNAs that are involved in the fine control of gene expression.

Aside from knowing the exact number of mRNA molecules in a human cell, which is currently being investigated using new sequencing technologies (de Klerk et al. 2014; Kellis et al. 2014), one of the great challenges of the next decade will be to decipher the posttranscriptional interactions between coding and ncRNAs in the control of gene expression.

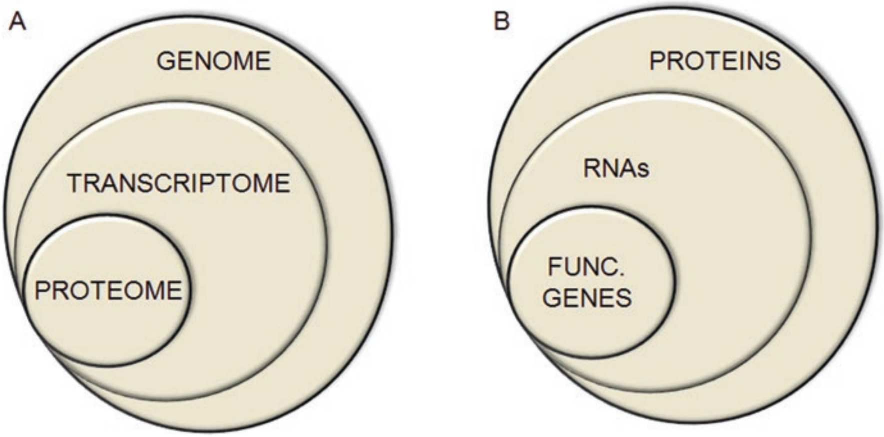


Fig. 1.5 Two ways to interpret the functioning the genome and the relative number of molecular entities. (**a**) In functional terms only a part of the genome encodes RNAs from which only a small fraction encodes proteins. (**b**) However, in numerical terms the set of functional genes transcribe a larger number of mRNAs from which a larger number of proteins is translated. The part **a** of this figure was conceived by Dr. Sven Diederichs (German Cancer Research Institute, DKFZ, Heidelberg, Germany) who allowed their use

In fact, the human genome was revealed to be more than just a collection of protein-coding genes and their splice variants, rather, it displays extensive antisense, overlapping, and ncRNA expression (Taft et al. 2010; https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet).

In mammals, the vast majority of the genome is transcribed into ncRNAs, which exceed the number of protein-coding genes (Liu and Taft 2013). These molecules are characterized by the absence of protein-coding capacity, but these RNAs have been described as key regulators of gene expression (Geisler and Coller 2013).

ncRNAs are grouped into two major classes based on their transcript size: small ncRNAs (19–30 nt) and long noncoding RNAs (200 nt to ~100 kilobases). These groups are distinct in their biological functions and mechanisms of gene regulation (Geisler and Coller 2013; Fatica and Bozzoni 2014; Neguembor et al. 2014).

Furthermore, ncRNAs can be grouped into a third class of housekeeping ncRNAs, which are normally constitutively expressed and include ribosomal (rRNAs), transfer (tRNAs), small nuclear (snRNAs), small nucleolar (snoRNAs), and regulatory noncoding RNAs (rnRNAs) (Ponting et al. 2009; Bratkovic and Rogelj 2014).

Small ncRNAs are primarily associated with the 5′ or 3′ regions of protein-coding genes, and based on their precursors and mechanism of action, they have been divided into three main classes: miRNAs, small interfering RNAs (siRNAs), and PIWI-associated RNAs (piRNAs). These ncRNAs are involved in posttranscriptional gene regulation through translational repression or RNAi (Sana et al. 2012).

Interestingly, the aberrant expression of small ncRNAs has been associated with a wide variety of human diseases, including cancer, central nervous system disorders, and cardiovascular diseases (Taft et al. 2010; Sana et al. 2012) (Table 1.1).

For much of the last decade, special attention has been paid to research into long noncoding RNAs (lncRNAs), as these molecules tend to be shorter and have fewer introns than protein-coding transcripts (Ravasi et al. 2006). lncRNAs are considered to be the most numerous and functionally diverse class of RNAs (Derrien et al. 2011). Over 15,000 lncRNAs have already been identified, and this number is constantly increasing (Kozlowska et al. 2021; Napoli 2021; Derrien et al. 2012; Fatica and Bozzoni 2014).

Amidst the great discoveries being made during this time of genome exploration, RNA is beginning to take center stage, and lncRNAs are a major part of this. These molecules are more abundant and functional than previously imagined, and they have been shown to be key players in gene regulation, genome stability, and chromatin modifications. Therefore, the identification and characterization of the function of lncRNAs have added a high degree of complexity to the comprehension of the structure, function, and evolution of our genome.

lncRNAs can be grouped into one or more of five categories based on their position relative to protein-coding genes: (1) sense or (2) antisense, when they overlap with one or more exons of another transcript on the same or opposite strand, respectively; (3) bidirectional, when the expression of a lncRNA and a neighboring coding transcript on the opposite strand is initiated in close genomic proximity; (4) intronic, when the lncRNA is fully derived from the intron of a second transcript; or (5)

Table 1.1 Major RNA classes found in eukaryotic cells

	Class	Symbol	Characteristics	References
Classical RNAs	Messenger RNAs	mRNAs	Variable in the size (average size about 2.2 kb) depending on the coded protein. Its linear structure includes a 5′ G cap, the 5′ UTR, AUG start codon, coding sequence (CDS), stop codon, 3′ UTR and the poly A tail. Account 1–2% of the total cellular RNA.	Roundtree et al. (2017)
	Transfer RNAs	tRNAs	tRNA was the first type of noncoding RNA to be characterized. This class of RNAs has cloverleaf secondary structure and has variable size ranging 70–100 nt. The residues 34, 35, and 36 are complementary to the mRNA codons located at CDS. For this reason they are considered as adaptors between mRNAs and elongation peptide chains.	Seal et al. (2020)
	Ribosomal RNAs	rRNAs	Components of ribosomes along with ribosomal proteins. The eukaryotic ribosome is denominated as the 80S ribosome and comprises two subunits: 60S (large subunit) and 40S (small subunit). The large subunit contains the high molecular weight 28S rRNA (~5000 nt) and two low molecular weight rRNAs, the 5.8 rRNA (156 nt) and 5.0S rRNA (121 nt), while the small subunit contains 18S RNA (1870 nt).	Seal et al. (2020)
Small noncoding RNAs	MicroRNAs	miRNAs	Drosha and Dicer dependent small ncRNAs; average size about 18–25 nt; account 1–2% of the human genome; more than 2500 miRNAs have been identified in the human genome to this date. They act in the post-transcriptional gene regulation of approximately 60% of human genes that encode proteins; guide suppression of translation.	Szilágyi et al. (2020) and Seal et al. (2020)
	Small interfering RNAs	siRNAs	Average size about 19–23 nt; made by Dicer processing; guide sequence specific degradation of target mRNA.	Nikam and Gore (2018)
	PIWI-interacting RNAs	piRNAs	piRNAs are animal-specific; size about 26–30 nt; bind PIWI-clade Argonaute proteins (PIWI proteins); dicer independent; transcribed from genomic loci denominated piRNA clusters; mostly restricted to the germline and somatic cells bordering the germline. The most elucidated piRNAs function is silencing transposons in the germ line through transcriptional and post-transcriptional mechanisms.	Szilágyi et al. (2020)
	Small nuclear RNAs	snRNA	Abundant transcripts that accumulate in the nucleus; typically range from 60 to 220 nt in length; accordingly on their sub-nuclear localization, structure, and function the snRNAs can be classified into one of three groups: spliceosomal U snRNA, small nucleolar RNA (snoRNA), and small Cajal body-specific RNA (scaRNA); snRNA associates with proteins to form small nuclear ribonucleoprotein particles (snRNPs); several snRNAs (U1, U2, U4, U5, and U6) are involved in pre-mRNA splicing of introns.	Guiro and Murphy (2017)

(continued)

Table 1.1 (continued)

	Class	Symbol	Characteristics	References
	Small nucleolar RNAs	snoRNA	Transcripts of around 60–170 nt, present at a lower copy number (~104 copies per cell for a single species); can be classified into three classes: C/D box snoRNAs (SNORDs), H/ACA box snoRNAs (SNORAs), and small Cajal body-specific RNAs (scaRNAs); generally involved in the modification of target ribosomal RNAs.	Seal et al. (2020) and Abel and Rederstorff (2019)
	Pyknons		Pyknons are specific human/primate-specific DNA motifs at least 16 nucleotides long that are repeated in blocks in intergenic and intronic regions; show variable lengths; are found more often in the 3′ untranslated regions (UTRs) of genes than in other regions of the genome.	Meynert and Birney (2006)
	Transcription initiation RNAs	tiRNAs	Located in the nucleus, average size about 18 nt; have the highest density just downstream of RNA polymerase II transcription start sites; show patterns of positional conservation; preferentially located in GC-rich promoters.	Tao et al. (2019)
	Centromere repeat associated short interacting RNA	crasiRNAs	Average size about 34–42 nt; processed from long dsRNAs; may play important function on eukaryotic centromere establishment.	Carone et al. (2013)
	Telomere-specific small RNAs	tel-sRNAs	Average size about 24 nt; Dicer independent; 2′-O-methylated at the 3′ terminus; evolutionarily conserved from protozoa to mammals; but have not been described in human up to now. Can play important roles in telomere structure and/or functions.	Cao et al. (2009)
	YRNAs (Ro-associateY)	YRNAs	Components of the Ro60 ribonucleoprotein particle; transcripts typically range 100–20 nt; secondary structures fold into a characteristic stem-loop; the human genome encodes four highly conserved types of YRNA transcripts: YRNA1 (RNY1), YRNA3 (RNY3), YRNA4 (RNY4), and YRNA5 (RNY5). The YRNAs are essential for initiation of DNA replication.	Szilágyi et al. (2020)
	Vault RNAs	vtRNA	Transcripts with lengths of 88–100 nt found in many eukaryotes; identified as components of ribonucleoprotein complexes known as vaults; the human genome contains four vault genes (VTRNA1-1, VTRNA1-2, VTRNA1-3, and VTRNA2-1). The function of vaults remains unclear.	Szilágyi et al. (2020) and Seal et al. (2020)
	RNA-glycan conjugates	GlycoRNAs	Located on the cell surface; conserved small noncoding RNAs bear sialylated glycans; present in multiple cell types and mammalian species; size ˂200 nt; may play a role in immune signaling.	Flynn et al. (2021)

Long noncoding RNAs	Circular RNAs and Circular intronic RNAs	circRNAs and CiRNAs	Noncoding RNAs mostly produced from protein-coding genes; characterized by a covalently closed-loop structure, generated by a process called backsplicing; circRNAs are produced from exonic sequence, while ciRNAs are produced from intronic sequence. They were shown acting as sponges for miRNAs, binding to RNA-binding protein (RBP), regulators of splicing and transcription.	Szilágyi et al. (2020) and Seal et al. (2020)
	Long intergenic ncRNAs	lincRNAs	ncRNA longer than 200 nucleotides that do not overlap annotated coding genes; most are localized in the nucleus; acts by remodeling chromatin and genome architecture, in the RNA stabilization, and in the transcriptional and post-transcriptional regulation.	Jarroux et al. (2017)
	Natural antisense	Nats	ncRNAs transcribed in the opposite direction of protein-coding genes; found in eukaryotes and prokaryotes; acts by modulations of chromatin changes, gene silencing, and RNA editing.	Zhao et al. (2018)
	TElomere repeat-containing RNA	TERRA	Mostly localize at telomeres; length ranges from 100 bp to 9 kb in mammalian cells; it has been suggested that TERRA plays an important player in telomere maintenance and genome stability.	Barral and Déjardin (2020)
	Transcribed ultraconserved regions	T-UCRs	Genomic regions conserved across orthologous regions of human, rat, and mouse; sequence longer than 200 bp; classified into: intergenic, intronic, exonic, partly exonic, or exonic containing. It has been suggested that T-UCRS are involved in cancer, regulating different mechanisms, such as: altered interactions with microRNA, direct binding to target mRNA, or hypermethylation of CpG island promoters.	Fabris and Calin (2017)

References

Ian A Roundtree, Molly E Evans, Tao Pan, Chuan He. **Dynamic RNA Modifications in Gene Expression Regulation.** Review Cell. 2017 Jun 15;169(7):1187–1200. doi: 10.1016/j.cell.2017.05.045**.**

Melinda Szilágyi, Ondrej Pös, Éva Márton, Gergely Buglyó, Beáta Soltész, Judit Keserű, András Penyige, Tomas Szemes, Bálint Nagy. **Circulating Cell-Free Nucleic Acids: Main Characteristics and Clinical Application.** Review Int J Mol Sci 2020 Sep 17;21(18):6827. doi: 10.3390/ijms21186827.

Xinyue Zhao, Jingrui Li, Bi Lian, Hanqing Gu, Yan Li, Yijun Qi. **Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA.** Nat Commun 2018 Nov 29;9(1):5056. doi: 10.1038/s41467-018-07500-7.

(continued)

Table 1.1 (continued)

Ruth L Seal, Ling-Ling Chen, Sam Griffiths-Jones, Todd M Lowe, Michael B Mathews, Dawn O'Reilly, Andrew J Pierce, Peter F Stadler, Igor Ulitsky, Sandra L Wolin, Elspeth A Bruford. **A guide to naming human non-coding RNA genes.** Review EMBO J. 2020 Mar 16;39(6):e103777. doi: 10.15252/embj.2019103777. Epub 2020 Feb 24.
Joana Guiro, **Shona Murphy**. **Regulation of expression of human RNA polymerase II-transcribed snRNA genes.** Review Open Biol. 2017 Jun;7(6):170073. doi: 10.1098/rsob.170073.
Yoann Abel, **Mathieu Rederstorff**. **SnoRNAs and the emerging class of sdRNAs: Multifaceted players in oncogenesis.** Review Biochimie. 2019 Sep;164:17–21. doi: 10.1016/j.biochi.2019.05.006. Epub 2019 May 9.
Alison Meynert, Ewan Birney. Picking pyknons out of the human genome. Review Cell 2006 Jun 2;125(5):836–8. doi: 10.1016/j.cell.2006.05.019.
Julien Jarroux, Antonin Morillon, Marina Pinskaya. History, Discovery, and Classification of lncRNAs. Review Adv Exp Med Biol. 2017;1008:1–46. doi: 10.1007/978-981-10-5203-3_1.
Rahul R Nikam, Kiran R Gore. **Journey of siRNA: Clinical Developments and Targeted Delivery.** Review Nucleic Acid Ther. 2018 Aug;28(4):209–224. doi: 10.1089/nat.2017.0715. Epub 2018 Mar 27.
En-Wei Tao, Wing Yin Cheng, Wei-Lin Li, Jun Yu, Qin-Yan Gao. **tiRNAs: A novel class of small noncoding RNAs that helps cells respond to stressors and plays roles in cancer progression**. Review J Cell Physiol 2020 Feb;235(2):683–690. doi: 10.1002/jcp.29057. Epub 2019 Jul 8.
Dawn M Carone, Chu Zhang, Laura E Hall, Craig Obergfell, Benjamin R Carone, Michael J O'Neill, Rachel J O'Neill. **Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading.** Chromosome Res. 2013 Mar;21(1):49–62. doi: 10.1007/s10577-013-9337-0. Epub 2013 Feb 8.
Amandine Barral, Jérome Déjardin. Telomeric Chromatin and TERRA. J Mol Biol. 2020 Jul 10;432(15):4244–4256. doi: 10.1016/j.jmb.2020.03.003. Epub 2020 Mar 7.
Linda Fabris, George A. Calin Understanding the Genomic Ultraconservations: T-UCRs and Cancer. Int Rev Cell Mol Biol. 2017;333:159–172. doi: 10.1016/bs.ircmb.2017.04.004. Epub 2017 Jun 7.
Ryan A Flynn, Kayvon Pedram, Stacy A Malaker, Pedro J Batista, Benjamin A H Smith, Alex G Johnson, Benson M George, Karim Majzoub, Peter W Villalta, Jan E Carette, Carolyn R Bertozzi. **Small RNAs are modified with N-glycans and displayed on the surface of living cells.**
Fang Cao, Xiangzhi Li, Samantha Hiew, Hugh Brady, Yifan Liu, Yali Dou. **Dicer independent small RNAs associate with telomeric heterochromatin.** RNA Jul;15(7):1274–81. doi: 10.1261/rna.1423309. Epub 2009 May 21.

intergenic, wherein an lncRNA is located within a gene (Poting et al. 2009). Most lncRNAs are transcribed by RNA Pol II and are often polyadenylated and have splice sites (Guttman et al. 2009; 2013; Mercer et al. 2013). However, they are devoid of obvious ORFs (Fatica and Bozzoni 2014).

The functional characterization of several mammalian regulatory lncRNAs has identified many biological roles, such as dosage compensation, genomic imprinting, cell cycle regulation, pluripotency, retrotransposon silencing, meiotic entry and telomerase length, and gene expression through chromatin modulation (Wery et al. 2011; Wilusz 2016; Nagano and Fraser 2011).

The number of lncRNAs with described functions is steadily increasing, and many of these reports revolve around the regulatory capacity of lncRNAs. These molecules localize both to the nucleus and to the cytosol and can act at virtually every level during gene expression (Batista and Chang 2013; Van et al. 2014). Nuclear lncRNAs act as modulators of protein-coding gene expression and can be subdivided into *cis*-acting RNAs, which act in proximity to their site of transcription, or *trans*-acting lncRNAs, which work at distant loci. Both *cis*- and *trans*-acting lncRNAs can activate or repress transcription via chromatin modulation (Penny et al. 1996; Pandey et al. 2008; Nagano et al. 2008; Chu et al. 2011; Plath et al. 2003; Bertani et al. 2011).

Cytoplasmic lncRNAs can modulate translational control via sequences that are complementary to transcripts that originate from either the same chromosomal locus or independent loci. Target recognition occurs through base pairing (Batista and Chang 2013).

RNA-Seq, the most powerful methodology for de novo sequence discovery, has been used to identify and analyze the expression of new lncRNAs in different cell types and tissues. Interestingly, sequencing experiments have shown that lncRNA expression is more cell-type specific than that of protein-coding genes (Derrien et al. 2012; Guttman et al. 2013; Mercer et al. 2008; Cabili et al. 2011; Pauli et al. 2012).

The identification of lncRNAs relies on the detection of transcription from genomic regions that are not annotated as protein coding. However, other similarly robust methodologies have been used in the identification of lncRNAs, including the following: (1) Tiling arrays: this technology enables the analysis of global transcription from a specific genomic region and was initially used to both identify and analyze the expression of lncRNAs; (2) serial analysis of gene expression (SAGE): this methodology allows both the quantification and the identification of new transcripts throughout the transcriptome; (3) cap analysis gene expression (CAGE): this methodology is based on the isolation and sequencing of short cDNA sequence tags that originate from the 5′ end of RNA transcripts; (4) chromatin immunoprecipitation (ChIP): this method allows the isolation of DNA sequences that are associated with a chromatin component of interest, thereby allowing the indirect identification of many unknown lncRNAs; and (5) RNA-Seq: in a single sequencing run, this methodology produces billions of reads that are subsequently aligned to a reference genome (Fatica and Bozzoni 2014).

Transcriptome research began in parallel with the genome project because of Craig Venter's idea to sequence the "most important" genes, i.e., the functioning genome. This directive clearly fell upon mRNAs, as this type of RNA carries the protein code. Of course, this concept has not changed and mRNAs are still of central importance; however, what followed was the subsequent discovery of a large number of different ncRNAs whose functions are linked to the fine control of gene expression, often controlling the translation of mRNAs into proteins, i.e., posttranscriptional control as it is exerted by miRNAs. In its broadest sense, the transcriptome is undoubtedly more complex than anyone previously imagined.

1.3 The Transcriptome and miRNome Are Closely Associated: The Role of microRNAs, a Class of Noncoding RNAs Linked to the Fine Control of Gene Expression

Cellular gene expression is governed by a complex, multi-faceted network of regulatory interactions. MicroRNAs (miRNAs) are recognized as important components that shape the transcriptome and also complement and extend the regulation that occurs in other levels. MiRNAs are small regulatory RNAs that produce a complex topology of gene expression that impacts cellular biological functions.

The discovery of the first miRNA in *Caenorhabditis elegans* (Lee et al. 1993; Wightman and Ruykun 1993) almost 30 years ago has led into a new era in molecular biology. Hundreds of different miRNAs have been identified in humans, many of which are conserved in other animals (Bartel 2008). MiRNAs play important roles in cellular homeostasis, development of organs and systems, cell differentiation, proliferation, metabolism, pluripotency, apoptosis, neural plasticity, memory, and others. They are also involved in pathologies, including metabolic disorders, cancers, neurodegenerative disorders, and infectious and autoimmune diseases (Bernstein et al. 2003; Ambros 2004; Neilson et al. 2007; Bushati and Cohen 2007; Stefani and Slack 2008; Baltimore et al. 2008; Bredy et al. 2011; Leonardo et al. 2012; Chen et al. 2016). MiRNAs are also promising candidates for new targeted therapeutic approaches and as biomarkers of disease. At approximately 22 nucleotides long, miRNAs are among the shortest known functional eukaryotic RNAs, and they repress most of the genes they regulate by just a small amount.

MicroRNAs are described as a group of endogenous small noncoding RNA of 18–22 nts in length. A large proportion of miRNAs are localized as cluster in the genome, predominantly distributed in intragenic and intergenic regions Kabekkodu et al. 2018). Most miRNAs genes are transcribed by RNA polymerase II as part of longer RNAs termed primary miRNAs (pri-miRNAs) (Lee et al. 2002, 2004; Cai et al. 2004). Although all canonical pri-miRNAs have a 5′ cap, they might not have the polyadenylation signal in the 3′ end (Ballarino et al. 2009).

Each pri-miRNA contains at least one region that folds back on itself to form a hairpin structure substrate for the microprocessor; a heterodimeric complex that contains one molecule of the endonuclease Drosha and two molecules of its dsRNA-binding partner DGCR8 (DiGeorge syndrome critical region gene 8) (Nguyen et al. 2015). Drosha has two RNase III domains that excised pri-miRNA hairpin which liberates the called "pre-miRNA." The resulting pre-miRNA consists of an approximately 70-nucleotide stem-loop structure characterized by imperfect base-pairing in the stem-loop and a two-nucleotide overhang at the 3′ end (Lee et al. 2002).

The pre-miRNA is subsequently exported to the cytoplasm by the nuclear transport protein exportin-5 (XPO5), in combination with the guanosine triphosphate (GTP) binding RAS-related nuclear protein (Ran-GTP) (Yi et al. 2003; Bohnsack et al. 2004; Lund et al. 2004; Okada et 2009). In the cytoplasm, the pre-miRNAs are further processed by the RNAse III enzyme Dicer liberating a 21–24 nt miRNA duplex. Several Dicer-associated proteins are known, including the double-stranded RNA-binding protein TRBP (TAR RNA-binding protein), PACT (protein activator of protein kinase R), and ADAR1 (adenosine deaminase acting on RNA) (Ha and Kim 2014; Hutvagner et al. 2001; Zhang et al. 2004).

After the sequential processing of the miRNA precursors, the miRNA duplex is loaded into an Argonaute (Ago) protein which is assisted by the HSP70-HSP90 chaperone machinery to form the RNA-induced silencing complex (RISC). Only one of the two strands of the miRNA duplex is retained in Ago proteins and stably forms RISC to mediate the recognition of the target mRNA (Chendrimada et al. 2005; Haase et al. 2005). The ratios of mature miRNAs derived from 5′(5p) and 3′(3p) sequences vary, and both strands of some miRNAs are functional (Chiang et al. 2010). Ago–miRNA complexes are guided to their specific targets through base pairing (Zealy et al. 2017) and perform its repression functions (Kawamata and Tomari 2010; Czech and Hannon 2011).

In addition, some miRNAs are produced by alternative pathways, independent of either Drosha- or Dicer-catalyzed cleavage by exploiting diverse RNases that normally catalyze the maturation of other types of transcripts (Yang and Lai 2010).

1.3.1 miRNA Regulatory Mechanisms

Besides numerous high-quality studies examining the biochemistry, biology, and genomics of miRNA-directed mRNA regulation, the factors that determine which mRNAs will be targeted, and the precise mechanisms of action remain incompletely understand. Extensive computational and experimental research over the last decade has substantially improved our knowledge of the mechanisms underlying miRNA-mediated gene regulation (Ameres and Zamore 2013; Yue et al. 2009; Ripoli et al. 2010; Bartel 2009; Chekulaeva and Filipowicz 2009; Brodersen and Voinnet 2009).

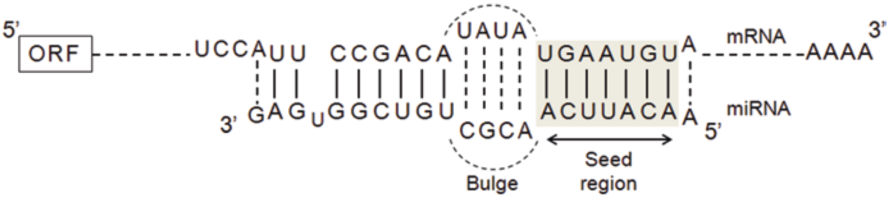


Fig. 1.6 Interaction of a miRNA with the 3′ UTR of its mRNA target by base pairing. (Figure adapted from Filipowicz et al. (2008) Nat Rev Genetics 9: 102–114)

In mammals, the dominate miRNA repression mode acts without slicing the target mRNA and does not require extensive pairing, as in plants. In turn, Ago proteins recruit downstream factors as the adaptor protein TNRC6 (Trinucleotide repeat containing 6) that interacts with the PABPC protein (poly(A)-binding) in the 3′ end of mRNA (Fig. 1.6). It recruits deadenylase complexes and most important CCR4-complex (Jonas and Izaurralde 2015). The deadenylases shorten the poly(A) tail and causes mRNA destabilization (Chen and Shyu 2011). CCR4-NOT complex recruits DDX6, a helicase that is reported to inhibit translation (Chu and Rana 2006; Jonas and Izarralde 2015). The repressive consequences of this TNRC6-mediated regulatory mode depend on the development context of the cell. Indeed, mRNA destabilization explains most of the repression mediated by mammalian miRNAs (Elchcorn et al. 2014). The miRNA-mediated RNA silencing processes are mainly localized in processing-bodies (P-bodies) in the cytoplasm (Liu et al. 2005).

For sites that promote TNCR6-mediated repression, miRNA recognized its target primarily through the pairing of its "seed" region that consists of an approximately 7-nucleotide domain at the 5′ end of the miRNA (Bartel 2009). These 7–8 nt sites mediate the bulk of the repression for each miRNA and are the sites identified by the most effective target-prediction tools (Bartel 2009; Agarwal et al. 2015). Pairing to the 3′ region of the miRNA can complement the seed region, but this supplementary pairing has little influence on efficacy (Wee et al. 2012; Salomon et al. 2015). Seed matches can occur in any region of an mRNA but are more likely to decrease mRNA expression when they are located in the 3′ untranslated region (3′ UTR) (Grimson et al. 2007; Forman et al. 2008, 2010; Gu et al. 2009). Because the region used to create the seed is so short, more than half of the protein-coding genes in mammals are regulated by miRNAs, and thousands of other mRNAs appear to have undergone negative selection to avoid seed matches with miRNAs that are present in the same cell (Baek et al. 2008; Lewis et al. 2003, 2005; Farh et al. 2005; Stark 2005; Lewis 2005).

The regulation of gene expression is a complicated multi-step process. Recently, multi-omics techniques were applied to determine the gene regulatory function of a given miRNA and revealed the influence of non-canonical regulations. Besides post-transcriptional regulation, miRNAs can also act on transcriptional regulation, including interaction with promoters, modulation of transcription factors, and interfering signal cascades as examples.

MiR-552-3p inhibit human cytochrome P450 2E1 (CYP2E1) via canonical mode. Interestingly, a non-seed sequence of miR-552 is complementary to CYP2E1 promoter inhibiting the binding of RNA polymerase II and leading to its silence. Mutations in the seed and non-seed regions of miR-552 show that this dual inhibition is more effective in gene regulation (Miao et al. 2016).

Several reports have demonstrated the network between miRNAs and transcription factors (Bracken et al. 2016). For example, miR-491-3p downregulated ABCB1 (a key factor in multidrug resistance) through canonical binding on its 3′ region, but it also targeted Sp3, a transcription factor of ABCB1. This dual inhibitory pathway increases the sensitivity of hepatoma cells to chemotherapeutic drugs (Zhao et al. 2017).

This miRNA dual inhibition also achieves key factors in signal cascades associated with the target gene. An example is miR-20a that downregulates CDKN1A expression through direct binding to the 3′ UTR and also interferes with the expression of factors in Smad/E2F-based repressor complex to indirectly reduce CDKN1A promoter activity (Sokolova et al. 2015). The advances in the field will increase our understanding about the complexity of regulatory miRNA functions and gene expression network.

Besides P-body, miRNAs can also traffic between various intracellular compartments (e.g., nucleus, mitochondria, Golgi and lysosome) contributing for the regulation of cellular functions. For example, miR-1 induced during myogenesis can enter mitochondria, where it stimulates translation of specific DNA-encoded transcripts (Sripada et al. 2012; Zhang et al. 2014; Chen et al. 2012).

In the nucleus, the first miRNA described was miR-21 in Hela cells (Meister et al. 2004). Further, several reports show the presence of miRISC components including Ago, Dicer, TRBP, and TRNC6/GW182 (Gagnon et al. 2014). Some results demonstrate that miRNA may have different nuclear functions as regulation of the noncoding RNA transcriptome, involvement in the cellular splicing program, and control of transcriptional gene activation (TGA) or transcriptional gene silencing (TGS) (Pu et al. 2019; Catalanotto et al. 2016). As a new and additional miRNA working mode, the exact mechanisms involved with their nuclear functions are not fully understood. In-depth studies are helping to gradually discover the inhibition or activation of intranuclear miRNAs.

1.3.2 Control of miRNA Expression

Advances in the biology of miRNAs have been revealing several regulatory mechanisms in the control of miRNAs biogenesis, maturation, and action in a cell-dependent manner under physiological and pathological conditions (Ha and Kim 2014; Treiber et al. 2017). MiRNA expression can be regulated at both transcriptional and post-transcriptional levels.

At DNA level, genetic alterations as amplification, mutations, and translocation in the genome can influence miRNA expression (Kabekkodu et al. 2018). Single nucleotide polymorphisms (SNPs) and epigenetic control of transcription, through classical mechanisms of acetylation/methylation of DNA/histones, also contribute for miRNA regulation (Pajares et al. 2021).

Once miRNAs are transcribed, modulation of Drosha expression and defects in exportin and Ago proteins also affect their expression (Romero-Cordoba et al. 2014; Ohtsuka et al. 2015; Gulyaeva and Kushlinskiy 2016). Although changes in accumulation and activity of both microprocessor and Dicer broadly affect miRNA production, individual miRNAs are differently sensitive, and some miRNAs expression can be impacted more than others. Adenosine to inosine (A-to-I) RNA editing of miRNA catalyzed by adenosine deaminase acting on RNA (ADAR) proteins may affect the stability, biogenesis, and target recognition of microRNA leading to changes in gene expression (Nishikura et al. 2013).

Biosynthesis and maturation of miRNAs can also be influenced by RNA-binding proteins (RBPs), which can interact with key enzymes such as DROSHA/DGCR8/DICER and the RISC complex (Ota et al. 2013). Examples of such mechanisms are illustrated by the stabilization of pri- and/or pre-miR-144 by BUD13 and Interleukin Enhancer Binding Factor 3 (ILF3) that lead to increased levels of mature forms. The use of proteomics-based pull-down approach and UV cross-linking followed by immunoprecipitation (eCLIP) analysis have expanded the list of RNA binding proteins (RBPs) that regulates miRNA biogenesis (Treiber et al. 2017; Nussbacher et al. 2018).

In addition, expression of other endogenous competing RNAs (ceRNAs), such as pseudogenes, circular RNAs, and long ncRNAs (lncRNAs), can act as "sponges" and impair specific miRNA–mRNA interactions (Thomson et al. 2016). Unexpectedly, some miRNAs are destabilized by specific interactions with mRNAs. This post-transcriptional regulation of miRNAs is called target-directed miRNA degradation (TDMD) and is mediated by transcripts containing sequences that have a near-perfect match with miRNAs and centered mismatches. Recent structural analyses of AGO2 and mutational analyses of miRNAs and their respective targets revealed that the shape of the AGO2 central cleft and the centered mismatches in the miRNA targets allow for modifications of the miRNA 3′ end by unknown enzymes. These modifications lead to 3′ end remodeling and eventually the decay of miRNAs (Park et al. 2019; Sheu-Gruttadauria et al. 2019).

1.3.3 Extracellular miRNAs

Additionally, to their well-known intracellular functions, miRNAs can also be secreted in extracellular complexes as extracellular vesicles (EVs) or associated with lipoproteins and ribonucleoproteins (Wagner et al. 2013; Arroyo et al. 2011; Valadi 2007). Circulating miRNAs have been studied in patient samples and animal models in the context of cancer, diabetes, cardiovascular disease, sepsis, and various other physiological and pathophysiological states (Cui et al. 2019; Villard et al. 2015; Cortez et al. 2011). Extracellular miRNAs have been identified in different body fluids (Wu et al. 2019; Kim et al. 2019; Mariner et al. 2018; Mall et al. 2013), and several findings indicate their utility as readily accessible biomarkers.

The demonstration that extracellular miRNAs could be transported from donor to recipient cells indicates a potential role as mediators of cell to cell communication. Some studies have shown that miRNAs are not randomly exported. Comparison of miRNA's expression levels in a variety of cells lines with their release exosomes indicates that a subset of miRNAs is preferentially selected (Guduric-Fuchs et al. 2012). Besides the considerable recent scientific advances in the field, the exact mechanisms involved in miRNA sorting, export, and uptake are still not fully understood (Vu et al. 2020).

For example, miR-223/105/106a are upregulated in HDL particles from patients with familial hypercholesterolemia. Cholesterol uptake could be decreased through the HDL-mediated miR-223 transport by targeting the scavenger receptor class B member 1 (SRB1) mRNA (SBR1 functions as a receptor for HDL) in cultured hepatocytes (Bayraktar et al. 2017). Nonsmall cell lung cancer cell lines secret EVs containing miR-21/29a that can bind to mouse toll-like receptor 7 (TLR7) and human TLR8 in tumor-associated macrophages, leading to nuclear factor κB (NF-κB) activation and secretion of the pro-metastatic inflammatory cytokines tumor necrosis factor α (TNF-α) and interleukin 6 (Fabbri 2018).

Another important feature of extracellular miRNAs that has gained lot of attention is their immunomodulatory roles. For example, EVs from regulatory T cells contain several miRNAs and miRNA precursor (Aiello et al. 2017; Okoye et al. 2014). The uptake of EVs containing miR-142 and miR150 by dendritic cells reduces the expression of cytokine IL-6 and increases the IL-10, interfering with antigen processing and presentation in these cells and inhibiting immune activation (Tung et al. 2020; Naqvi et al. 2016).

1.3.4 An Example of the Biological Consequence of miRNAs: Their Role in Immune Diseases

It is clear that as important modulators of cell differentiation, proliferation, and survival, miRNAs contribute to various diseases at the molecular levels. A study conducted in 2016 determined the miRNAs profile in human tissue biopsies from different organs and revealed a tissue-specific expression (Ludwig et al. 2016). Alterations in miRNAs expression pattern have been demonstrated in a variety of human diseases, such as cancer, autoimmunity, cardiovascular diseases, and viral infections, confirmed as a casual factor in disease progression

(Paul et al. 2018). For example, in systemic lupus erythematosus, upregulation of several miRNAs was described to be involved in disease progression. DNMT1 expression was decreased by upregulation of miR148a and miR-21 leading to a DNA hypomethylation pattern (Wang et al. 2018). MiRNAs are also associated with SLE disease activity index (SLEDAI) and are a good predictor of disease activity (Khoshmirsafa et al. 2019).

In rheumatoid arthritis (Donate et al. 2013), studies showed the overexpression of miRNA-155, miR-124a, and miR-223 in various tissue and immune cells, such as CD4$^+$ T cells and CD14$^+$ cells derived from synovial fluid, B cells, macrophages, and PBMCs. The association between miR-155 expression in PBMCs and swollen joints indicates that miR-155 promotes the progression of RA by triggering production and recruitment of cytokines (Zakeri et al. 2019; Tavasolian et al. 2018).

In multiple sclerosis**,** EAE murine models show upregulation of let-7e in CD4$^+$ T cells, amplifying the function of Th1 and Th17 cells and increasing IL10 activity, leading to an increase in the severity of EAE. Furthermore, miR-155 has been shown to be able to regulate Th17 cells by controlling the suppressive effects of Jarid2, which functions as a recruiter of PRC2. Decreased expression of miR-320a has been found in B cells, leading to overexpression of MMP-9 (Yang et al. 2018).

Due to their roles in disease progression, there are increasing interests in the development of new miRNA-target therapies. Interestingly, one of the first miRNA-based molecules to enter clinical development was the LNA miravirsen, a 15-nucleotide antisense RNA oligo with complementarity to the 5′ end of miR-122, for the treatment of HCV (Van Rooij et al. 2014, Elmen et al. 2008a, b). Since then considerable progress was reached in terms of therapeutic approaches (Rupaimoole et al. 2017).

1.4 Conclusion

Early on, transcriptome research was intertwined with the genome. Much of this was due to the mapping of ESTs, and sequencing dominated the scene. Through the use of EST clones and the application of technical concepts such as nucleic acid hybridization, researchers began to use arrayed filters to explore the transcriptional expression of a large number of genes in a single experiment.

The constant improvement of these DNA arrays led to the fabrication of high-density arrays and, finally, microarrays.

At the same time, sequencing also underwent significant changes involving automation and the endless quest to increase the number of reads, and this contributed substantially to a better understanding of the diversity of the transcriptome. Indeed, transcriptome research was rooted in these two major technological approaches (i.e., large-scale hybridization and sequencing).

What made microarrays robust and increased their popularity was the increase in the number of sequences deposited on the slides (currently, these slides contain the entire human or mouse functional genome), the sensitivity of the method (currently, experiments are being performed with nanogram amounts of total RNA to screen the entire functional genome), the simplicity of its use, its commercial availability, and the availability of bioinformatics packages dedicated to analyzing the large amounts of data being generated.

Of key importance was the development of statistical procedures for the analysis of large amounts of data, which opened the door for biostatisticians and bioinformaticians.

All of these ongoing technological advances have contributed to the consolidation of the concept of the transcriptome. Unlike the genome, which is essentially static, the transcriptome is variable and is dependent on normal physiological, pathological, or environmental conditions. Moreover, it is composed of not only mRNAs but also noncoding RNAs, including miRNAs.

This concept has provided the opportunity for all types of biomedical research to re-examine their results in light of transcriptomics.

Acknowledgments Our laboratories are funded by the following agencies: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, São Paulo, Brazil), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasília, Brazil), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brasília, Brazil, through financial code 001), Institut National de la Santé et de la Recherche Médicale (INSERM, Paris, France), and ARCUS-PACA (Provence-Alpes-Côte d'Azur)-Brésil Cooperation Agreement.

References

Adams J (2008) Sequencing human genome: the contributions of Francis Collins and Craig Venter. Nat Educ 1(1):133

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC (1992) Sequence identification of 2375 human brain genes. Nature 355:632–634

Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC (1993a) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nat Genet 4:373–338

Adams MD, Kerlavage AR, Fields C, Venter JC (1993b) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nat Genet 4:256–267

Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. elife 4:e05005

Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R (2010) Generating and navigating proteome maps using mass spectrometry. Nat Rev Mol Cell Biol 11:789–801

Aiello S et al (2017) Extracellular vesicles derived from T regulatory cells 78. suppress T cell proliferation and prolong allograft survival. Sci Rep 7:11518

Alfaro JA, Bohländer P, Dai M, Filius M, Howard CJ, van Kooten XF, Ohayon S, Pomorski A, Schmid S, Aksimentiev A, Anslyn EV, Bedran G, Cao C, Chinappi M, Coyaud E, Dekker C, Dittmar G, Drachman N, Eelkema R, Goodlett D, Hentz S, Kalathiya U, Kelleher NL, Kelly

RT, Kelman Z, Kim SH, Kuster B, Rodriguez-Larrea D, Lindsay S, Maglia G, Marcotte EM, Marino JP, Masselon C, Mayer M, Samaras P, Sarthak K, Sepiashvili L, Stein D, Wanunu M, Wilhelm M, Yin P, Meller A, Joo C (2021) The emerging landscape of single-molecule protein sequencing technologies. Nat Methods 18:604–617

Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7:55–65

Altelaar AF, Munoz J, Heck AJ (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 14:35–48

Ambros V (2004) The functions of animal microRNAs. Nature 431:350–355

Ameres SL, Zamore PD (2013) Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol 14(8):475–488

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106

Anders S, Pyl PT, Huber W (2015) HTSeq – a Python framework to work with high- throughput sequencing data. Bioinformatics 31:166–169

Anderson L (2014) Six decades searching for meaning in the proteome. J Proteome 107:24–30

Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF, Pogosova-Agadjanyan EL et al (2011) Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. Proc Natl Acad Sci U S A 108:5003–5008

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185:405–416

Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP (2008) The impact of microRNAs on protein output. Nature 455:64–71

Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R et al (2005) The external RNA controls consortium: a progress report. Nat Methods 2:731–734

Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, Cavalieri D, Gaasterland T, Hingamp P et al (2002) Microarray gene expression data (MGED) society. Standards for microarray data. Science 298:539

Ballarino M, Pagano F, Girardi E, Morlando M et al (2009) Coupled RNA processing and transcription of intergenic primary microRNAs. Mol Cell Biol 29:5632–5638

Baltimore D, Boldin MP, O’Connell RM, Rao DS, Taganov KD (2008) MicroRNAs: new regulators of immune cell development and function. Nat Immunol 9:839–845

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233

Batista PJ, Chang HY (2013) Long noncoding RNAs: cellular address codes in development and disease. Cell 152:1298–1307

Bayraktar R, Van Roosbroeck K, Calin GA (2017) Cell-to-cell communication: MicroRNAs as hormones. Mol Oncol 11(12):1673–1686

Bernard K, Auphan N, Granjeaud S, Victorero G, Schmitt-Verhulst AM, Jordan BR, Nguyen C (1996) Multiplex messenger assay: simultaneous, quantitative measurement of expression for many genes in the context of T cell activation. Nucleic Acids Res 24:1435–1443

Bernstein E, Kim SY, Carmell MA et al (2003) Dicer is essential for mouse development. Nat Genet 35:215–217

Bertani S, Sauer S, Bolotin E et al (2011) The noncoding RNA mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol Cell 43:1040–1046

Bertucci F, Bernard K, Loriod B, Chang YC, Granjeaud S, Birnbaum D, Nguyen C, Peck K, Jordan BR (1999) Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for smalls samples. Hum Mol Genet 9:1715–1722

Bohnsack MT, Czaplinski K, Gorlich D (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA 10:185–191

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2):185–193

Botwell D (1999) Options available -from start to finish- for obtaining expression data by microarray. Nat Genet 21:2–32

Bourgon R, Gentleman R, Huber W (2010) Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A 107:9546–9551

Bracken CP, Scott HS, Goodall GJ (2016) A network-biology perspective of microRNA function and dysfunction in cancer. Nat Rev Genet 17(12):719–732

Bratkovic T, Rogelj B (2014) The many faces of small nucleolar RNAs. Biochim Biophys Acta 1839:438–443

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA et al (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. Nat Genet 29:365–371

Bredy TW, Lin Q, Wei W, Baker-Andresen D, Mattick JS (2011) MicroRNA regulation of neural plasticity and memory. Neurobiol Learn Mem 96:89–94

Breitling R, Herzyk P (2005) Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. J Bioinforma Comput Biol 3:1171–1189

Brenner S, Jacob F, Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature 190:576–581

Brodersen P, Voinnet O (2009) Revisiting the principles of microRNA target recognition and mode of action. Nat Rev Mol Cell Biol 10(2):141–1488

Bushati N, Cohen SM (2007) MicroRNA functions. Annu Rev Cell Dev Biol 23:175–205

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25:1915–1927

Cahan P, Rovegno F, Mooney D, Newman JC, St. Laurent G III, McCaffrey TA (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene 401:12–18

Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are pro- cessed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10:1957–1966

Camargo AA, Samaia HP, Dias-Neto E, Simão DF, Migotto IA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S et al (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc Natl Acad Sci U S A 98:12103–12108

Cantor CR (1990) Orchestrating the human genome project. Science 248:49–51

Catalanotto C, Cogoni C, Zardo G (2016) MicroRNA in control of gene expression: an overview of nuclear functions. Int J Mol Sci 17(10):1712

Chatterjee A, Ahn A, Rodger EJ et al (2018) A guide for designing and analyzing RNA-Seq data. In: Raghavachari N, Garcia-Reyero N (eds) Gene expression analysis. Methods in molecular biology, vol 1783. Humana Press, New York

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP (1996) Accessing genetic information with high-density DNA arrays. Science 274:610–614

Chekulaeva M, Filipowicz W (2009) Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. Curr Opin Cell Biol 21:452–460

Chen CY, Shyu AB (2011) Mechanisms of deadenylation-dependent decay. Wiley Interdisc Rev RNA 2:167–183

Chen WS, Leung CM, Pan HW, Hu LY, Li SC, Ho MR, Tsai KW (2012) Silencing of miR-1-1 and miR-133a-2 cluster expression by DNA hypermethylation in colorectal cancer. Oncol Rep 28:1069–1076

Chen Y, Dougherty ER, Bittner ML (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Opt 2:364–374

Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF et al (1998) Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324

Chen J-Q, Papp G, Szodoray P, Zeher M (2016) The role of microRNAs in the pathogenesis of autoimmune diseases. Autoimmun Rev 15(12):1171–1180

Chendrimada TP, Gregory RI, Kumaraswamy E et al (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature 436:740–744

Chiang HR, Schoenfeld LW, Ruby JG et al (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev 24:992–1009

Chu C, Qu K, Zhong FL et al (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. Mol Cell 44:667–678

Chu CY, Rana TM (2006) Translation repression in human cells by microRNA-induced gene silencing requires RCK/p54. PLoS Biol 4:e210

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet 32:490–495

Clément-Ziza M, Gentien D, Lyonnet S, Thiery JP, Besmond C, Decraene C (2009) Evaluation of methods for amplification of picogram amounts of total RNA for whole genome expression profiling. BMC Genomics 26(10):246

Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, Xiao W, Laudanski K, Brownstein BH, Elson CM et al (2005) Application of genome-wide expression analysis to human health and disease. PNAS 102(13):4801–4806

Conesa A, Madrigal P, Tarazona S et al (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13

Cortez MA, Bueso-Ramos C, Ferdin J (2011) MicroRNAs in body fluids – the mix of hormones and biomarkers. Nat Rev Clin Oncol 8:467–477

Cui M, Wang H, Yao X et al (2019) Circulating microRNAs in cancer: potential and challenge. Front Genet 10:626

Czech B, Hannon GJ (2011) Small RNA sorting: matchmaking for argonautes. Nat Rev Genet 12:19–31

de Klerk E, den Dunnen JT, t Hoen PA (2014) RNA sequencing: from tag-based profiling to resolving complete transcript structure. Cell Mol Life Sci 71(18):3537–3551

Degrelle SA, Hennequet-Antier C, Chiapello H, Piot-Kaminski K, Piumi F, Robin S, Renard JP, Hue I (2008) Amplification biases: possible differences among deviating gene expressions. BMC Genomics 9:46

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4:P3

Derrien T, Guigo R, Johnson R (2012) The long non-coding RNAs: a new (p)layer in the "dark matter". Front Genet 2:107

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D et al (2013) A comprehensive evaluation of normalization methods for illumine high-throughput RNA sequencing data analysis. Brief Bioinform 14(6):671–683

Djebali S, Davis CA, Merkel A et al (2012) Landscape of transcription in human cells. Nature 489:101–108

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21

Donate PB, Fornari TA, Macedo C, Cunha TM, Nascimento DC, Sakamoto-Hojo ET, Donadi EA, Cunha FQ, Passos GA (2013) T cell post-transcriptional miRNA-mRNA interaction networks identify targets associated with susceptibility/resistance to collagen-induced arthritis. PLoS One 8(1):e54803

Duewer DL, Jones WD, Reid LH, Salit M (2009) Learning from microarray interlaboratory studies: measures of precision for gene expression. BMC Genomics 10:153

Dufva M (2005) Fabrication of high quality microarrays. Biomol Eng 22:173–184

Dujon B (1998) European functional analysis network (EUROFAN) and the functional analysis of the Saccharomyces cerevisiae genome. Electrophoresis 19:617–624

Edwards D (2003) Non-linear normalization and background correction in onechannel cDNA microarrays studies. Bioinformatics 19:825–833

Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA et al (2014) mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Mol Cell 56:104–115

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. PNAS 95(25):14863–14868

Elmen J et al (2008a) Antagonism of microRNA-122 in mice by systemically administered LNA-antimiR leads to up-regulation of a large set of predicted target mRNAs in the liver. Nucleic Acids Res 36:1153–1162

Elmen J et al (2008b) LNA-mediated microRNA silencing in non-human primates. Nature 452:896–899

Epstein JR, Leung AP, Lee KH, Walt DR (2003) High-density, microsphere based fiber optic DNA microarrays. Biosen Bioeletron 18:541–546

Fabbri M (2018) MicroRNAs and miRceptors: a new mechanism of action for intercellular communication. Philos Trans R Soc Lond Ser B Biol Sci 373(1737):20160486

Fang Z, Cui X (2010) Design and validation issues in RNA-seq experiments. Brief Bioinform 12(3):280–287

Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP (2005) The widespread impact of mammalian microRNAs on mRNA repression and evolution. Science 310:1817–1821

Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15:7–21

Ferguson JA, Steemers FJ, Walt DR (2000) High-density fiber optic DNA random microsphere array. Anal Chem 72:5618–5624

Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of posttranscriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9(2):102–114

Fisher RA (1935) The design of experiments. Oliver & Boyd, Oxford, England, p 251

Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28(24):3169–3177

Foreman RE, George AL, Reimann F, Gribble FM, Kay RG (2021) Peptidomics: a review of clinical applications and methodologies. J Proteome Res, July 16. (Epub ahead of print. PMID: 34270237)

Forler S, Klein O, Klose J (2014) Individualized proteomics. J Proteome 107C:56–61

Forman JJ, Legesse-Miller A, Coller HA (2008) A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci U S A 105:14879–14884

Gagnon KT, Li L, Chu Y, Janowski BA, Corey DR (2014) RNAi factors are present and active in human cell nuclei. Cell Rep 6(1):211–221

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for trasncriptome annotation and quantification using RNA-seq. Nat Methods 8:469–477

Geisler S, Coller J (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol 14:699–672

Gentleman RC, Carey VJ, Bates DM (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10):80.1–80.16

Gentleman RC, Carey VJ, Huber W, et al (2005) Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York, 473 p

Gershon D (2002) Microarray technology, an array of opportunities; technology feature. Nature 416:885–891

Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86

Granjeaud S, Nguyen C, Rocha D, Luton R, Jordan BR (1996) From hybridization image to numerical values: a practical, high throughput quantification system for high density filter hybridizations. Genet Anal Biomol Eng 12:151–162

Granjeaud S, Bertucci F, Jordan BR (1999) Expression profiling: DNA arrays in many guises. BioEssays 21:781–790

Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H (1992) Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. Mamm Genome 3:609–661

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27:91–105

Gu S, Jin L, Zhang F, Sarnow P, Kay MA (2009) Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. Nat Struct Mol Biol 16:144–150

Guduric-Fuchs J, O'Connor A, Camp B et al (2012) Selective extracellular vesicle-mediated export of an overlapping set of microRNAs from multiple cell types. BMC Genomics 13:357

Gulyaeva LF, Kushlinskiy NE (2016) Regulatory mechanisms of microRNA expression. J Transl Med 14:143

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E et al (2004) Decoding randomly ordered DNA arrays. Genome Res 14:870–877

Guo Y, Ye F, Sheng Q, Clark T, Samuels DC (2013) Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. https://doi.org/10.1093/bib/bbt069

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 154:240–251

Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15:509–524

Haase AD, Jaskiewicz L, Zhang H, Laine S et al (2005) TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. EMBO Rep 6:961–967

Hänzelmann S, Castelo R, Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 14:7

Heber S, Sick B (2006) Quality assessment of Affymetrix GeneChip data. OMICS 10(3):358–368

Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18(suppl 1):S96–S104

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Bioinformatics 4(2):249–264

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J et al (2005) Multiple-laboratory comparison of microarray platforms. Nat Methods 2:345–350

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356

Järvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O (2004) Are data from different gene expression microarray platforms comparable? Genomics 83:1164–1168

Jonas S, Izaurralde E (2015) Towards a molecular understanding of microRNA-mediated gene silencing. Nat Rev Genet 16:421–433

Jordan BR (1998) Large scale expression measurement by hybridization methods: from high-density membranes to "DNA chips". J Biochem 124:251–258

Jordan B (2012) The microarray paradigm and its various implementations. In: Jordan B (ed) Microarrays in diagnostics and biomarker development. Current and Future Applications. Springer, Berlin/Heidelberg

Joyce S, Ternette N (2021) Know thy immune self & non-self: proteomics informs on the expanse of self and non-self, and how and where they arise. Proteomics 26:e2000143

Kabekkodu SP, Shukla V, Varghese VK, D'Souza J et al (2018) Clustered miRNAs and their role in biological functions and diseases. Biol Rev Camb Philos Soc 93(4):1955–1986

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Kawamata T, Tomari Y (2010) Making RISC. Trends Biochem Sci 35:368–376

Kellis M, Wold B, Snyder MP et al (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A 111:6131–6138

Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2:183–201

Khoshmirsafa M et al (2019) Elevated expression of miR-21 and miR-155 in peripheral blood mononuclear cells as potential biomarkers for lupus nephritis. Int J Rheum Dis 22:458–467

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36

Kim YJ, Yeon Y, Lee WJ et al (2019) Comparison of MicroRNA expression in tears of normal subjects and Sjögren syndrome patients. Invest Ophthalmol Vis Sci 60:4889–4895

Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T (2002) Improved background correction for spotted DNA microarrays. J Comp Biol 9:55–66

Kozlowska J, Kolenda T, Poter P, Sobocińska J, Guglas K, Stasiak M, Bliźniak R, Teresiak A, Lamperska K (2021) Long intergenic non-coding RNAs in HNSCC: from "Junk DNA" to important prognostic factor. Cancers (Basel) 13(12):2949

Lamarre S et al (2018) Optimization of an RNA- Seq differential gene expression analysis depending on biological replicate number and library size. Front Plant Sci 9:108

Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854

Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization. EMBO J 21:4663–4670

Lee Y, Kim M, Han J, Yeom KH et al (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23:4051–4060

Leonardo TR, Schultheisz HL, Loring JF, Laurent LC (2012) The functions of micro-RNAs in pluripotency and reprogramming. Nat Cell Biol 14:1114–1121

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. Cell 115:787–798

Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15–20

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323

Liu G, Mattick JS, Taft RJ (2013) A meta-analysis of the genomic and transcriptomics composition of complex life. Cell Cycle 12:2061–2072

Liu J, Valencia-Sanchez MA, Hannon GJ, Parker R (2005) Microrna-dependent localization of targeted mRNAs to mammalian p-bodies. Nat Cell Biol 7:719–723

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680

Lönnstedt I, Speed T (2002) Replicated microarray data. Stat Sin 12:31–46

Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T et al (2016) Distribution of miRNA expression across human tissues. Nucleic Acids Res 44(8):3865–3877

Lund JM, Alexopoulou L, Sato A, Karow M, Adams NC, Gale NW, Iwasaki A, Flavell RA (2004) Recognition of single-stranded RNA viruses by toll-like receptor 7. Proc Natl Acad Sci U S A 101:5598–5603

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ (2009) GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinf 10:161

Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, Shi T, Tong W, Shi L et al (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenom J 10:278–291

Maeda N, Kasukawa T, Oyama R et al (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet 2:e62

Mall C, Rocke DM, Durbin-Johnson B, Weiss RH (2013) Stability of miRNA in human urine supports its biomarker potential. Biomark Med 7:623–631

MAQC Consortium (2006) The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24:1151–1161

Mariner PD, Korst A, Karimpour-Fard A, Stauffer BL, Miyamoto SD, Sucharov CC (2018) Improved detection of circulating miRNAs in Serum and plasma following rapid heat/freeze cycling. Microrna 7(2):138–147

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12:671–682

Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell 15(2):185–197

Mercer TR, Mattick JS (2013) Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Biol 20:300–307

Miao L, Yao H, Li C, Pu M et al (2016) A dual inhibition: microRNA-552 suppresses both transcription and translation of cytochrome P450 2E1. Biochim Biophys Acta 1859(4):650–662

Minnier J, Pennock ND, Guo Q et al (2018) RNA-Seq and expression arrays: selection guidelines for genome-wide expression profiling. In: Raghavachari N, Garcia-Reyero N (eds) Gene expression analysis. Methods in molecular biology, vol 1783. Humana Press, NY, New York

Moorcroft MJ, Meuleman WR, Latham SG, Nicholls TJ, Egeland RD, Edwin M, Southern EM (2005) In situ oligonucleotide synthesis on poly(dimethylsiloxane): a flexible substrate for microarray fabrication. Nucleic Acids Res 33:e75

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian trasncriptome by RNA-Seq. Nat Methods 5(7):621–628

Nagano T, Fraser P (2011) No-nonsense functions for long noncoding RNAs. Cell 145:178–181

Nagano T, Mitchell JA, Sanz LA et al (2008) The air noncoding RNA epigenetically silencestranscription by targeting G9a to chromatin. Science 322:1717–1720

Napoli S (2021) LncRNAs and Available Databases. Methods Mol Biol 2348:3–26

Naqvi AR, Fordham JB, Ganesh B, Nares S (2016) MiR-24, miR-30b and miR-142-3p interfere with antigen processing and presentation by primary macrophages and dendritic cells. Sci Rep 6:1–12

Neguembor MV, Jothi M, Gabellini D (2014) Long noncoding RNAs, emerging players in muscle differentiation and disease. Skelet Muscle 4:8

Neilson JR, Zheng GX, Burge CB, Sharp PA (2007) Dynamic regulation of miRNA expression in ordered stages of cellular development. Genes Dev 21:578–589

Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, Jordan BR (1995) Differential gene expression inthe murine thymus assayed by quantitative hybridization of arrayed cDNA clones. Genomics 29:207–216

Nguyen TA, Jo MH, Choi YG, Park J et al (2015) Functional anatomy of the human microprocessor. Cell 161:1374–1387

Nishikura K, Sakurai M, Ariyoshi K, Ota H (2013) Antagonistic and stimulative roles of ADAR1 in RNA silencing. RNA Biol 10:1240–1247

Nussbacher JK, Yeo GW (2018) Systematic discovery of RNA binding proteins that regulate MicroRNA levels. Mol Cell 69:1005–1016.e7

Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP et al (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12:1749–1755

Nygaard VL, Hovig E (2006) Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling. Nucleic Acids Res 34:996–1014

Ohtsuka M, Ling H, Doki Y, Mori M, Calin G (2015) MicroRNA processing and human Cancer. J Clin Med 4:1651–1667

Okoye IS et al (2014) MicroRNA-containing T-regulatory-cell-derived exosomes suppress pathogenic T helper 1 cells. Immunity 41:89–103

Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet 2:173–179

Oshlack A, Robinson MD, Young M (2010) From RNA-seq reads to differential expression results. Genome Biol 11:220–230

Ota H, Sakurai M, Gupta R et al (2013) ADAR1 forms a complex with Dicer to promote microRNA processing and RNA-induced gene silencing. Cell 153:575–589

Padron G, Domont GB (2014) Two decades of proteomics in Latin America: a personal view. J Proteome 107C:83–92

Pajares MJ, Alemany-Cosme E, Goñi S, Bandres E, Palanca-Ballester C, Sandoval J (2021) Epigenetic regulation of microRNAs in cancer: shortening the distance from bench to bedside. Int J Mol Sci 22:7350

Pandey RR, Mondal T, Mohammad F et al (2008) Kcnq1ot1antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32:232–246

Park J, Seo JW, Ahn N, Park S, Hwang J, Nam JW (2019) UPF1/SMG7-dependent microRNA-mediated gene regulation. Nat Commun 10:4181

Patel RK, Jain M (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619

Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, Fang H, Kawasaki ES, Hager J et al (2006) Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol 24(9):1140–1150

Paul P, Chakraborty A, Sarkar D, Langthasa M et al (2018) Interplay between miRNAs and human diseases. J Cell Physiol 233(3):2007–2018

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22:577–591

Penny GD, Kay GF, Sheardown SA et al (1996) Requirement for Xist in X chromosome inactivation. Nature 379:131–137

Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, Mariage-Samson R, Houlgatte R, Soularue P, Auffray C (1996) Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 6:492–503

Pietu G, Mariage-Samson R, Fayein NA, Matingou C, Eveno E, Houlgatte R, Decraene C, Vandenbrouck Y, Tahi F et al (1999) The genexpress image knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. Genome Res 9:195–209

Plath K, Fang J, Mlynarczyk-Evans SK et al (2003) Role of histone H3 lysine 27 methylation in X inactivation. Science 300:131–135

Ploner A, Miller LD, Hall P, Bergh J, Pawitan Y (2005) Correlation test to assess low-level processing of high-density oligonucleotide microarray data. BMC Bioinf 6:80

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–641

Pu M, Chen J, Tao Z, Miao L, Qi X, Wang Y, Ren J (2019) Regulatory network of miRNA on its target: coordination between transcriptional and post-transcriptional regulation of gene expression. Cell Mol Life Sci 76(3):441–451

Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet 2:418–427

Rapaport F, Khanin R, Liang Y et al (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 14:3158

Ravasi T, Suzuki H, Pang KC et al (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res 16:11–19

Ripoli A, Rainaldi G, Rizzo M, Mercatanti A, Pitto L (2010) The fuzzy logic of microRNA regulation: a key to control cell complexity. Curr Genomics 11:350–353

Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK (2007) A comparison of background corrections methods for two-color microarrays. Bioinformatics 23(20):2700–2707

Rocha D, Carrier A, Naspetti M, Victorero G, Anderson E, Botcherby M, Nguyen C, Naquet P, Jordan BR (1997) Modulation of mRNA levels in the presence of thymocytes and genome mapping for a set of genes expressed in mouse thymic epithelial cells. Immunogenetics 46:142–151

Romero-Cordoba SL, Salido-Guadarrama I, Rodriguez-Dorantes M, Hidalgo-Miranda A (2014) miRNA biogenesis: biological impact in the development of cancer. Cancer Biol The 15:1444–1455

Rupaimoole R, Slack FJ (2017) MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. Nat Rev Drug Discov 16:203–222

Salomon WE, Jolly SM, Moore MJ, Zamore PD, Serebrov V (2015) Single-molecule imaging reveals that Argonaute reshapes the binding properties of its nucleic acid guides. Cell 162:84–95

Sana J, Faltejskova P, Svoboda M, Slaby O (2012) Novel classes of non-coding RNAs and cancer. J Transl Med 10:103–123

Schena M, Shanon D, Heller R et al (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A 93:10614–10619

Seyednasrollah F, Laiho A, Elo LL (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 16(1):59–70

Sheu-Gruttadauria J, Pawlica P, Klum SM, Wang S et al (2019) Structural basis for target-directed microRNA degradation. Mol Cell 75:1243–1255.e7

Shi L, Reid LH, Jones WD et al (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24(9):1151–1161

Shi L, Campbell G, Jones WD et al (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28(8):827–838

Shin C, Nam JW, Farh KK et al (2010) Expanding the microRNA targeting code: functional sites with centered pairing. Mol Cell 38:789–802

Silver JD, Ritchie ME, Smyth GK (2009) Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. Biostatistics 10(2):352–363

Singh RL, Maganti RJ, Jabba SV, Wang M, Deng G, Heath JD, Kurn N, Wangemann P (2005) Microarray-based comparison of three amplification methods for nanogram amounts of total RNA. Am J Phys Cell Physiol 288:C1179–C1189

Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol 10:974–978

Slonim DK, Yanai I (2009) Getting started in gene expression microarray analysis. PLoS Comput Biol 5(10):e1000543

Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (eds) Bioinformatics and computational biology solutions using R and bioconductor. Springer, New York, pp 397–420

Sogayar MC, Camargo AA, Bettoni F et al (2004) A transcript finishing initiative for closing gaps in the human transcriptome. Genome Res 14:1413–1423

Sokolova V, Fiorino A, Zoni E, Crippa E, Reid JF, Gariboldi M, Pierotti MA (2015) The effects of miR-20a on p21: two mechanisms blocking growth arrest in TGF-beta-responsive colon carcinoma. J Cell Physiol 230(12):3105–3114

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinf 14:91–108

Sripada L, Tomar D, Prajapati P, Singh R, Singh AK, Singh R (2012) Systematic analysis of small RNAs associated with human mitochondria by deep sequencing: detailed analysis of mitochondrial associated miRNA. PLoS One 7(9):e44873

Stark A, Brennecke J, Bushati N et al (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3Í′ UTR evolution. Cell 123:1133–1146

Stefani G, Slack FJ (2008) Small non-coding RNAs in animal development. Nat Rev Mol Cell Biol 9:219–230

Stekel D (2003) Microarray bioinformatics. Cambridge University Press, Cambridge. ISBN: 9780521525879

Strausberg RL, Riggins GL (2001) Navigating the human transcriptome. Proc Natl Acad Sci U S A 98:11837–11838

Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102:15545–15550

Sudo K, Chinen K, Nakamura Y (1994) 2058 expressed sequence tags (ESTs) from a human fetal lung cDNA library. Genomics 24:276–279

Sudo H, Mizoguchi A, Kawauchi J, Akiyama H, Takizawa S (2012) Use of non-amplified RNA samples for microarray analysis of gene expression. PLoS One 7:e31397

Taft RJ, Pang KC, Mercer TR et al (2010) Non-coding RNAs: regulators of disease. J Pathol 220:126–139

Takeda J, Yano H, Eng S, Zeng Y, Bell GI (1993) Construction of a normalized directionally cloned cDNA library from adult heart and analysis of 3040 clones by partial sequencing. Hum Mol Genet 2:1793–1798

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 12:2213–2223

Tavasolian F, Abdollahi E, Rezaei R, Momtazi-Borojeni AA et al (2018) Altered expression of microRNAs in rheumatoid arthritis. J Cell Biochem 119(1):478–487

The ENCODE Consortium (2011) Standards, guidelines and best practices for RNA-Seq. Available at http://encodeproject.org/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

Thomson DW, Dinger ME (2016) Endogenous microRNA sponges: evidence and controversy. Nat Rev Genet 17:272–283

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc 7:562–578. (Erratum in: Nat Protoc 2014 9: 2513)

Treiber T, Treiber N, Plessmann U, Harlander S et al (2017) A compendium of RNA-binding proteins that regulate microRNA biogenesis. Mol Cell 66:270–284

Tung SL et al (2020) Regulatory T cell extracellular vesicles modify T-effector 77. cell cytokine production and protect against human skin allograft damage. Front Cell Dev Biol 8:317

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98(9):5116–5121

Valadi H, Ekstrom K, Bossios A et al (2007) Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9:654–659

Van Dijk EL, Jaszczyszyn Y, Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res 322:12–20

Van Haaften RI, Schroen B, Janssen BJ, van Erk A, Debets JJ, Smeets HJ, Smits JF, van den Wijngaard A, Pinto YM, Evelo CT (2006) Biologically relevant effects of mRNA amplification on gene expression profiles. BMC Bioinf 7:200

Van Heesch S, Van Iterson M, Jacobi J et al (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol 15:R6

Van Rooij E, Sutherland LB, Qi X et al (2007) Control of stress-dependent cardiac growth and gene expression by a microRNA. Science 316:575–579

Velculescu VE, Zhang L, Zhou W et al (1997) Characterization of the yeast transcriptome. Cell 88:243–251

Villard A, Marchand L, Thivolet C, Rome S (2015) Diagnostic value of cell-free circulating microRNAs for obesity and type 2 diabetes: a meta-analysis. J Mol Biomark Diagn 6:251

Vu LT, Gong J, Pham TT, Kim Y, Le MTN (2020) microRNA exchange via extracellular vesicles in cancer. Cell Prolif 53(11):e12877

Wagner J, Riwanto M, Besler C et al (2013) Characterization of levels and cellular transfer of circulating lipoprotein-bound microRNAs. Arterioscler Thromb Vasc Biol 33:1392–1400

Wang X, Cairns MJ (2013) Gene set enrichment analysis of RNA-Seq data:integrating differential expression and splicing. BMC Bioinf 14(Suppl 5):S16

Wang J, Hu L, Hamilton SR, Coombes KR, Zhang W (2003) RNA amplification strategies for cDNA microarray experiments. Biotechniques 34:394–400

Wang Z, Heid B, Dai R, Ahmed SA (2018) Similar dysregulation of lupus-associated miRNAs in peripheral blood mononuclear cells and splenic lymphocytes in MRL/lpr mice. Lupus Sci Med 5:e000290

Watson JD (1990) The human genome project: past, present, and future. Science 248:44–49

Wee LM, Flores-Jasso CF, Salomon WE, Zamore PD (2012) Argo- naute divides its RNA guide into domains with distinct functions and RNA- binding properties. Cell 151:1055–1067

Wery M, Kwapisz M, Morillon A (2011) Noncoding RNAs in gene regulation. Wiley Interdiscip Rev Syst Biol Med 3:728–738

Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin´14 by lin‘4 mediates temporal pattern formation in C. elegans. Cell 75:855–862

Wilusz JE (2016) Long noncoding RNAs: re-writing dogmas of RNA processing and stability. Biochim Biophys Acta 1859:128–138

Wreschner DH, Herzberg M (1984) A new blotting medium for the simple Isolation and Identification of highly resolved messenger RNA. Nucleic Acids Res 12:1349–1359

Wu L, Zheng K, Yan C et al (2019) Genome-wide study of salivary microRNAs as potential noninvasive biomarkers for detection of nasopharyngeal carcinoma. BMC Cancer 19:843

Yang JS, Lai EC (2011) Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. Mol Cell 43:892–903

Yang X, Wu Y, Zhang B, Ni B (2018) Noncoding RNAs in multiple sclerosis. Clin Epigenetics 10

Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 17:3011–3016

Young MD, Wakefield MJ, Smyth GK et al (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14

Yue D, Liu H, Huang Y (2009) Survey of computational algorithms for microRNA target prediction. Curr Genomics 10:478–492

Zakeri Z, Salmaninejad A, Hosseini N, Shahbakhsh Y et al (2019) MicroRNA and exosome: key players in rheumatoid arthritis. J Cell Biochem 2019

Zealy RW, Wrenn SP, Davila S, Min KW, Yoon JH (2017) MicroRNA-binding proteins: specificity and function. Wiley Interdisc Rev RNA 8:5

Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W (2004) Single processing center models for human Dicer and bacterial RNase III. Cell 118:57–68

Zhang X, Zuo X, Yang B et al (2014) MicroRNA directly enhances mitochondrial translation during muscle differentiation. Cell 158(3):607–619

Zhao N, Hashida H, Takahashi N, Misumi Y, Sakaki Y (1995) High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. Gene 156:207–213

Zhao Y, Qi X, Chen J et al (2017) The miR-491-3p/Sp3/ABCB1 axis attenuates multidrug resistance of hepatocellular carcinoma. Cancer Lett 408:102–111

Chapter 2
Alternative Splicing of Pre-messenger RNA

Vanessa Cristina Arfelli and Leticia Fröhlich Archangelo

2.1 Splicing and the Splicing Machinery

Since the discovery that eukaryotic genes are discontinuous, much has been learnt about how its transcriptional products are processed to generate the protein-coding mRNAs (Sharp 1994, 2005). Splicing is the cellular process in eukaryotic cells in which the intronic non-coding sequences are removed from the precursor mRNA (pre-mRNA) and exonic sequences are juxtaposed to yield mature functional mRNAs. This process is accomplished by a large machinery, the spliceosome, comprised of five small nuclear ribonucleoprotein particles (U1, U2, U3, U4, and U6 snRNP) and approximately 170 associated proteins, which combinational composition varies from one stage to the next throughout spliceosome cycle (Wahl et al. 2009).

The spliceosome assembly onto the pre-RNA is a highly dynamic process, where a series of consecutive steps produce the complexes E, A, B, C, P and ILS, respectively. The formation of these complexes is based on the establishment and dismantlement of several weak interactions between RNA:RNA, protein:RNA, and protein:protein molecules that act synergically to recognize and assemble onto the pre-RNA splice sites and to form the catalytically active structure (Wahl et al. 2009). The entire process is highly orchestrated and subject to different levels of regulation to guarantee the correct processing of mRNAs and the fidelity of the cellular transcriptome.

Three common consensus sequences define intronic regions of a pre-mRNA and are needed for the initial recognition and assembly by the spliceosome on the splice sites: the 5′ donor splice site (5′-ss) and the 3′ acceptor splice site (3′-ss) at both ends of the intron and a branchpoint sequence (BPS). Every intron almost invariantly contains GU dinucleotide at the 5′ end and AG dinucleotide at the 3′ end. These

V. C. Arfelli · L. F. Archangelo (✉)
Department of Cellular and Molecular Biology and Pathogenic Bioagents, Ribeirão Preto Medical School, University of São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil
e-mail: leticiafa@fmrp.usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_2

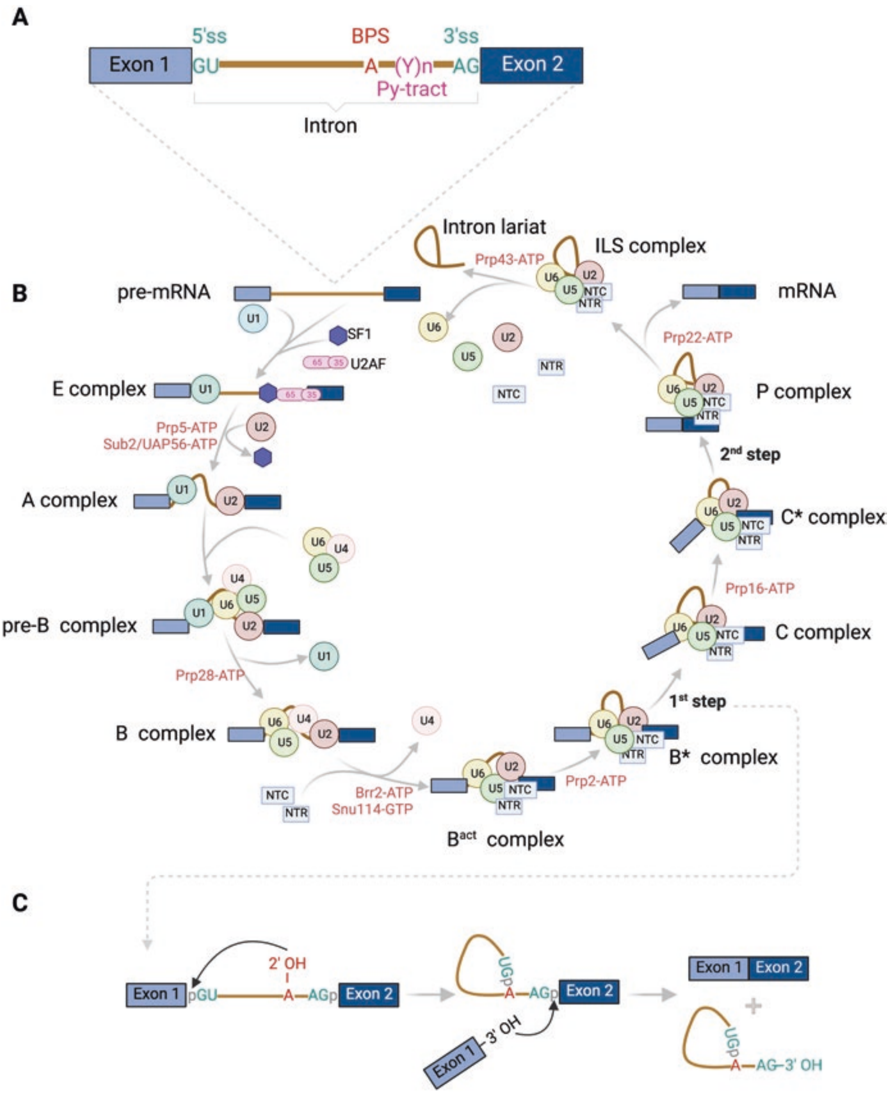


Fig. 2.1 *The splicing process*. (**a**) pre-mRNA structure: Exons 1 and 2 are represented in light and dark blue, respectively. The intron is represented by brown line. The 5′ donor splice site (5′-ss) is defined by the nucleotides GU, while the 3′ acceptor splice site (3′-ss) is defined by the nucleotides AG. In humans, the conserved branch point sequence (BPS) is yUnAy, which is represented in the scheme by the adenine (A). The poly-pyrimidine tract (Py-tract) is represented by a Y. (**b**) Splicing events coordinated by the major spliceosome: the stepwise binding of U2AF (pink), SF1 (purple), snRNPs (colored circles), and NTR and NTC (light blue) is depicted, as well as the action of helicases (red) to form the complexes E, A, B, C, P, and ILS, resulting in the mature mRNA and the removal of introns. (**c**) Transesterification reactions. The adenine (A) 2-OH in BPS mediates the nucleophilic attack to the phosphate group (p) in the 5′-ss of the intron. In the second transesterification reaction, the free 3′-OH of the first exon attacks the phosphate group in the 3′-ss, yielding junction of exons 1 and 2 and removal of the intron in a lariat form. (Made in ©BioRender – biorender.com)

dinucleotides are positioned inside longer consensus sequences known to influence the strength of the splicing sites. The BPS is located in the proximity upstream of the 3′-ss. A pyrimidine enriched sequence, known as polypyrimidine tract (Py-tract), lies between the BPS and the 3′-ss (Fig. 2.1a) (Reed 1996).

The stepwise assembly of the spliceosome begins with the Early complex (complex E) formation when U1 snRNP, through RNA–RNA base-pairing interactions, recognizes the 5′-ss. It is followed by ligation of splicing factor 1 (SF1/BBP) to BPS and its interaction with the U2 auxiliary factor large subunit (U2AF65), which in turn associates with the Py-tract, whereas the small subunit of the U2AF heterodimer (U2AF35) binds to the 3′-ss.

In the subsequent ATP-dependent step, the U2 snRNP is recruited to replace SF1 from its interaction with the BPS. It interacts with U1 snRNP and turns the E complex into a pre-spliceosome complex (complex A). The next step involves binding of the pre-assembled U4/U6 and U5 snRNPs to form the pre-B complex. The exit of U1 snRNP marks the formation of B complex. Although all the snRNPs are present at this point, the complex B is catalytically inactive. In order to activate the spliceosome, the complex has to go through a series of conformational and compositional changes turning it into an activated complex (complex B^{act} and further B*). Specifically, 5′-ss and 3′-ss are brought into proximity, U4/U6 duplex unwind, and the U4 snRNP is released, allowing U6 to interact with the 5′-ss.

Moreover, the NineTeen complex (NTC) and the NTC-related complex (NTR) are recruited. The activated B complex engages in the first catalytic reaction, additional rearrangements occur and generate the catalytic complex C (C* complex), which undergoes the second catalytic step of splicing. In the P complex, interactions between the three conserved elements of the intron (3′-ss, 5′-ss, and lariat junction) occur. Rearrangements promote mature mRNA release, leaving only the intron lariat spliceosome (ILS). Finally, the U2, U5, U6 snRNPs and NTC and NTR complexes are released to engage in an additional round of splicing and the post-spliceosome complex disassemble (Fig. 2.1b) (Matera and Wang 2014; Wahl et al. 2009; Wan et al. 2019).

A large amount of energy is devoted to RNA remodeling throughout spliceosome formation, which is employed by the action of numerous evolutionarily conserved DExD/H type RNA-dependent ATPases/helicases that act at specific steps of the splicing cycle to catalyze RNA–RNA rearrangements and RNP remodeling events (Cordin and Beggs 2013; Staley and Guthrie 1998).

Essentially, the splicing process of intron removal and ligation of the flanking exons entails two trans-esterification reactions involving functional groups in the 5′-ss, 3′-ss, and BPS regions of an intron. First, a nucleophilic attack by the 2′ hydroxyl group of a conserved adenosine residue within the BPS cleaves the phosphodiester bond within the 5′ exon–intron junction. The reaction generates a free 3′ hydroxyl group on the 5′ exon and a lariat intron intermediate. In the second reaction, the phosphodiester bond in the 3′ intron–exon junction is attacked by the 3′ hydroxyl group of the 5′ exon, displacing the lariat and promoting exons ligation (Padgett et al. 1986) (Fig. 2.1c).

The major spliceosome, built by the U1, U2, U4/U6, and U5 snRNPs particles, as described above, is responsible for removing the vast majority of pre-mRNA introns.

However, a distinct but structurally and functionally analogous spliceosome complex mediates the excision of a rare subset of evolutionary conserved introns that exhibit non-canonical consensus sequence, referred to as minor-class introns (Hall and Padgett 1994). The minor-class intron spliceosomes are low abundant and formed by the distinct but functionally analogous snRNPs U11, U12, U4atac, and U6atac together with the U5 snRNPs, which is a particle shared by both machineries. The much less frequent minor-class introns coexist with neighboring canonical major-class introns in a gene. The two spliceosome machineries undergo comparable dynamic rearrangements, with the main differences occurring at the early stages of intron recognition rather than during catalysis (Patel and Steitz 2003). The minor-class splicing follows the same two-step reactions and formation of a lariat intermediate as the major splicing. U11 base-pair with the characteristic longer and constrained consensus sequence at the 5′-ss of the minor class introns, whereas U12 base-pair with the BPS. The secondary structure of U11 and U12 mimics that of U1 and U2 snRNAs, respectively. Minor-class introns lack the Py-tract. Analogous to the major pathway, the U4atac snRNP chaperone U6atac into the spliceosome, preventing U6atac interaction with U12 and the 5′-ss before their helicase-dependent unwinding. Upon unwinding, U4atac is released, followed by rearrangements that permit catalytic activation (Patel and Steitz 2003).

Initial recognition and pairing of the 5′ and 3′ splice sites depend very much on the size of the intron and the distance between the splice sites affects the efficiency in which spliceosome assembles (Fox-Walsh et al. 2005). Because splice sites are recognized across an optimal nucleotide length, depending on the size of the intron or the flanking exons, the splice sites are recognized across the intronic or exonic segments, known as exon and intron definition models (Berget 1995; De Conti et al. 2013). When exons are small and introns are long, the splicing machinery forms across exons, whereas in genome architecture, where exons are large and introns are small, such as observed in lower eukaryotes, the intron definition prevails. In the human genome, where the majority of exons are short and introns are long, it is likely that the vast majority of splice sites are recognized across the exon (Fig. 2.2). The same splicing complexes formed across exon operate on intron definition in terms of composition and structure, and both exon and intron definition models may co-occur within the same pre-mRNA (De Conti et al. 2013; Li et al. 2019).

2.2 Alternative Splicing

Whereas some exons are constitutively spliced, that is, they are present in every mRNA produced from a given pre-mRNA, others are alternatively spliced to generate variable forms of mRNA from a single pre-mRNA. By shifting the exon usage, alternative splicing (AS) notably enhances the transcriptome and cellular proteome (Pan et al. 2008). Alternatively, spliced gene products may have related, distinct, or even opposing functions as well as non-functional properties, which may lead to the regulation of its expression through nonsense-mediated mRNA decay (NMD) or nuclear sequestration and turnover. Consequently, AS represents an important level of

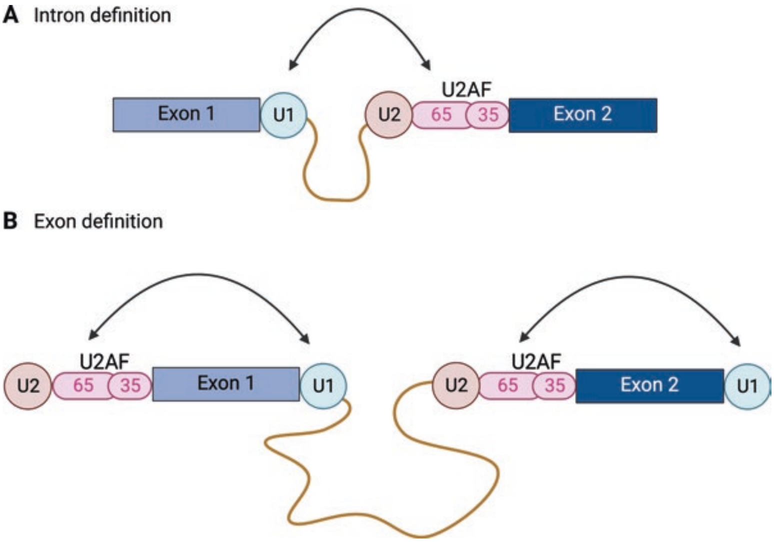


Fig. 2.2 *Intron and exon definition models*. (**a**) Intron definition model: the U1 and U2 snRNPs and U2AF interact to allow the pairing of the splice sites across an intron when the intron is short (<250 bp). (**b**) Exon definition model: the pairing of the splice sites occurs across the exon when separated by a long intron (>250 bp). (Made in ©BioRender – biorender.com)

regulation in gene expression and plays a critical role in biological processes such as development, cell differentiation, and response to environmental cues. AS frequency increases with species complexity and it has been proposed to be a driver of phenotypic complexity evolution in mammals. Accordingly, significative higher frequencies of AS events are observed in brain tissues throughout vertebrate species where regulation of these splicing events has been associated with evolutionary changes contributing to nervous system development (Barbosa-Morais et al. 2012; Merkin et al. 2012).

Virtually all human multi-exon genes undergo alternative splicing (Pan et al. 2008; Wang et al. 2008). The main types of AS events are exon skipping, mutually exclusive exon, alternative 5′ and 3′ splice site selection, and alternatively retained introns (Nilsen and Graveley 2010). Exon skipping takes place when a particular cassette exon is spliced out of the mature message. In a mutually exclusive exon AS event, one out of two exonic regions of a pre-mRNA is included in the final transcript when the other is excluded, and vice versa, so that these exons never coexist in the same product. Alternative splice site selection occurs when spliceosome recognizes and pairs with cryptic splice sites, resulting in the alternative splice site cleavage on the nearby exon. The fifth type of AS event, the retained intron, is the process by which a particular intron remains unspliced in the final transcript, and as a result, it triggers the downregulation of the transcript through NMD or nuclear sequestration and turnover (Braunschweig et al. 2014). In addition to the common types of AS, alternative promoter usage and alternative cleavage and polyadenylation yield different types of alternative transcript events, such as alternative first exon (AFE), tandem 3′ untranslated region (UTR), and alternative last exon (ALE).

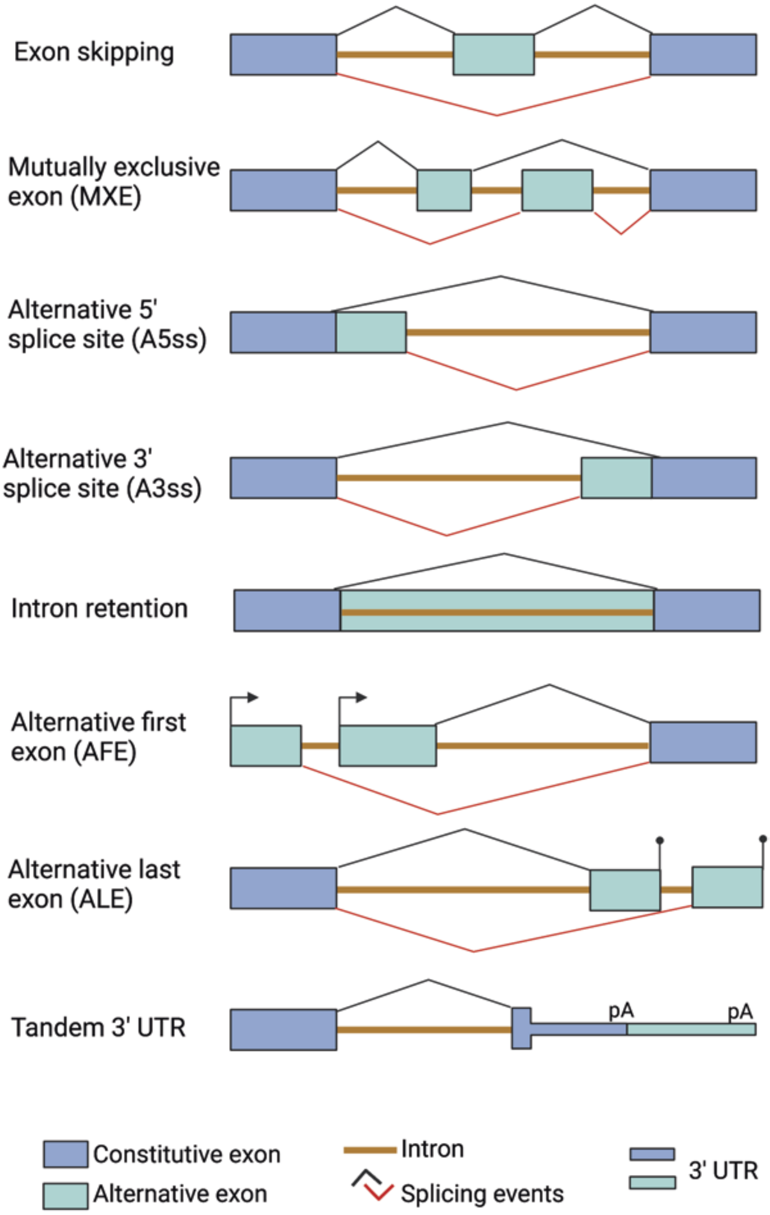

Exon skipping
Mutually exclusive exon (MXE)
Alternative 5' splice site (A5ss)
Alternative 3' splice site (A3ss)
Intron retention
Alternative first exon (AFE)
Alternative last exon (ALE)
Tandem 3' UTR
pA
pA
Constitutive exon
Alternative exon
Intron
Splicing events
3' UTR

Fig. 2.3 *Types of alternative splicing (AS) events*. (Made in ©BioRender – biorender.com)

In the first case, alternative promoter usage gives rise to mRNA isoforms with distinct 5′ UTR, and in the second case, the usage of alternative polyadenylation sites gives rise to transcripts with shorter or longer 3′ UTR and with distinct terminal exons (Wang et al. 2008) (Fig. 2.3). A combination of different modes of AS is often observed throughout a single precursor mRNA of multiexon genes.

In humans, the splice site sequences are highly degenerated (Sheth et al. 2006) and often not sufficient to define exon–intron boundaries. In addition, sequences that match the short consensus splice site signals are commonly found throughout the introns. In order to help the splice site selection, exons and their nearby intronic regions contain a variety of additional splicing regulatory elements (SREs). If they enhance exon inclusion, these elements are called exon splicing enhancers (ESE) or intronic splicing enhancers (ISE), depending on whether they are present in exonic or intronic regions. If they tend to repress exon inclusion, these elements are called exon or intron splicing silencers (ESS or ISS, respectively) (Zhang et al. 2008).

These cis-acting sequences within the pre-mRNA influence splicing through the binding of specific RNA-binding non-spliceosomal regulatory proteins, which either promote or hinder the spliceosome activity on the adjacent splicing sites (Cartegni et al. 2002).

The requirement for additional cis-acting and trans-acting elements to stabilize and target specific sites introduces another layer of complexity in the regulation of the splicing machinery and provides an important window for variations and diversity.

Two major classes of widely expressed trans-acting factors, namely the SR proteins and the heterogeneous ribonucleoproteins (hnRNPs), are involved in recognizing and binding the cis-elements within the RNA (Long and Caceres 2009; Martinez-Contreras et al. 2007).

Proteins of the SR family contain one or two N-terminal RNA recognition motif (RRM), which mediate binding to RNA and a C-terminal arginine-serine-rich (RS) domain, involved mainly in protein–protein interaction. They play an important role as general splicing factors and as regulators of alternative splicing (Manley and Krainer 2010). The hnRNPs form a group of structurally diverse RNA binding proteins involved in different stages of the RNA metabolism (Geuens et al. 2016).

Several reports describe the antagonistic function of SR and hnRNP proteins on alternative splicing (Cáceres et al. 1994). Typically, splicing enhancers are recognized by a member of the SR family, whereas hnRNP recognizes splicing silencers (Cartegni et al. 2002). SR sequence motifs are enriched in exonic sequence (Liu et al. 1998). When SR proteins are bound to ESEs they favor exon inclusion and prevent exon skipping. They can also promote exon definition by directly recruiting the splicing machinery through their RS domain and antagonizing nearby silencer elements.

On the other hand, hnRNP sequence motifs are enriched in introns. hnRNP represses splicing by directly antagonizing the recognition of splice sites or interfering with the binding of proteins bound to enhancers. Various hnRNPs regulate alternative splicing by stimulating exon skipping or intron retention (Fig. 2.4). To add an additional layer of complexity to this network, some SR protein can be implicated in splicing silencing when associated with introns, whereas some hnRNP can inhibit splicing from exonic locations. Specifically, SR and hnRNP protein activities differ depending on their position relative to the regulated splice, in which some SR protein can repress splicing, whereas hnRNP can enhance splicing depending on their position relative to the regulated site (Erkelenz et al. 2013; Matera and Wang 2014).

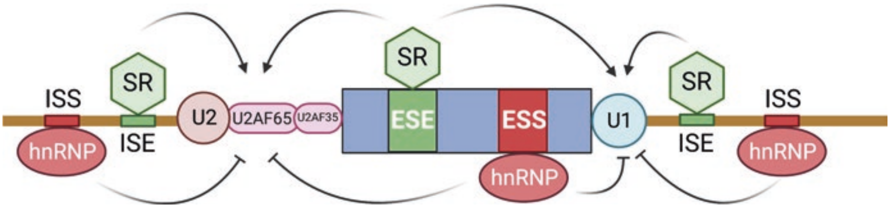


Fig. 2.4 *The antagonistic role of SR proteins and hnRNPs in splicing*. In general, when bound to enhancer sequences, SR proteins positively influence splicing promoting exon inclusion, exon definition, and recruiting the splicing machinery. hnRNPs, conversely, bind to silencing elements, inhibiting the recognition of splice sites, and interfering with the binding of proteins to enhancers. ESE: exonic splicing enhancer; ESS: exonic splicing silencer; ISE: intronic splicing enhancer; ISS: intronic splicing silencer. (Made in ©BioRender – biorender.com)

In the end, the decision of whether a specific site is selected or if a particular exon or intron is included or excluded from the final transcript is defined by the combinatorial interplay of positive and negative regulatory signals present in the RNA, the ultimate complexes formed by the trans-acting factors assembled on these regulatory sequences and how they influence the splicing machinery on the nearby splice sites. On top of that, variations in the relative concentrations of the antagonistic trans-acting elements may affect splice-site choice by tipping the balance in favor of different outcomes. Thus, the ratios of these antagonistic factors are likely to define a cellular code for establishing cell-specific patterns of splicing in multiple genes (Smith and Valcárcel 2000).

2.3 Mechanisms of Alternative Splicing Regulation

A much higher level of complexity is added to alternative splicing regulation when considering the fact that most splicing evens occur co-transcriptionally. The coupling of splicing and transcription implies a tight integration of alternative splicing with other gene regulatory pathways (Braunschweig et al. 2013).

2.3.1 RNA Architecture and Secondary Structures

RNA architecture impacts splicing outcomes. Secondary structures on the pre-mRNA can influence the accessibility of splice sites or cis-acting elements. These secondary structures are formed by intramolecular base-pairing and impact splice site selection positively or negatively. Thus, RNA folding may be regarded as an important component of AS regulation (McManus and Graveley 2011; Warf and Berglund 2010).

The mechanisms by which these structures influence splicing may involve local- and long-range interactions within the RNA molecules. Local pairing forms RNA structures that may prevent the binding of appropriate regulatory factors to the single-stranded molecule, which can happen when secondary structures are formed onto cis-regulatory elements or in its vicinity. Similarly, when secondary structures overlap splicing signals, such as 5′-ss, 3′-ss, BPS, and Py-tract, it can hamper recognition and assembly of spliceosome complexes onto these sites.

Conversely, long-range interactions place distant sequences into proximity, which promotes the looping out of specific regions of the pre-mRNA. These regions may contain cassette exons, a stretch of exonic and intronic sequences, or functionally active cis-acting elements that are drawn to alternative splicing regulation. Besides, approximation of distant regulatory elements and cognate factors to target exons may favor alternative splice site selection. The role of alternative competing RNA folding in the choice of alternative splice sites and mutually exclusive exons was first described for the drosophila *Dscam* gene, but similar modes of regulation have been proposed for human genes (Pervouchine et al. 2012).

While RNA secondary structure impacts splicing, its formation is condition-dependent and may also be subject to regulation. For instance, RNA helicases can unwind these structures and consequently regulate pre-mRNA splicing. Also, transcription rate can influence the folding of the RNA during synthesis and ultimately if slow-folding structures have enough time or not to assemble on the nascent pre-mRNA before splicing occurs.

2.3.2 Coupling Transcription to Alternative Splicing

Besides the interaction of snRNPs and splicing regulatory factors with pre-mRNA, the high-fidelity process of splice site selection also requires transcriptional machinery as well as chromatin modifiers and remodelers (Luco et al. 2011). Each step of transcription, namely initiation, elongation, and termination, contributes to how the nascent pre-mRNA is processed.

The impact of promoter usage in alternative splicing was first demonstrated on experimental models in which an artificial minigene under the regulation of different promoters exhibited a different pattern of alternative transcript. This led to the conclusion that promoters may contribute to AS by recruiting different molecules to the transcriptional complex, which in turn participate in the splicing regulation. The primary implication of these findings was that cell-specific AS may not simply result from the differential abundance of SR proteins but also from a more complex process involving cell-specific promoter occupation. Unlike the minigene experimental model, most of the genes are regulated by a single promoter in nature. In this case, the differential occupancy of the promoters by a variety of transcription factors and co-activators impacts AS. The promoter itself is responsible for recruiting splicing regulator factors to the site of transcription, possibly through the interaction with transcription factors bound to the promoter or transcriptional enhancers. Also, some of the effects of promoters on pre-mRNA splicing are mediated by proteins that function as dual transcription and splicing factors (Kornblihtt 2005).

The RNA polymerase II (RNAPII) has a major impact on alternative splicing, and its largest subunit C-terminal domain (CTD) plays a central role in coupling the two processes. In mammals, the CTD domain comprises 52 heptad repeats (YS_2PTS_5PS), subject to extensive phosphorylation. Phosphorylated CTD serves as a binding module for multiple mRNA processing factors, including splicing factors. Phosphorylation of the CTD repeats on serine 5 residues are essential for capping enzyme recruitment, whereas phosphorylation on serine 2 facilitates recruitment of cleavage and polyadenylation factors at 3′ ends of the RNA. Particularly, serine 2 phosphorylation was shown to be essential for the integration of transcription and splicing. In addition, the mediator complex, known to facilitate the interaction between the transcription pre-initiation complex on promoters with distant transcription regulatory factors bound to enhancers, also contacts splicing factors (David and Manley 2011).

The regulation of the RNAPII elongation rate constitutes another mechanism in which transcription affects AS. The RNAPII elongation rate is governed by the rates of RNA synthesis and translocation of the enzyme interspersed by acceleration, deceleration, backtracking, pausing and release, and sometimes with premature termination, which may occur while the RNAPII transcription elongation complex travels along a given gene. The use of RNAPII rate mutants to investigate the impact of elongation rate on genome-wide alternative splicing in human cells demonstrated that both slow and fast transcription changed the alternative splicing of thousands of exons. Slower transcription rates mainly contribute to cassette exons' inclusion, whereas a faster transcription leads to their skipping from mature mRNAs (Saldi et al. 2016).

Two models were proposed to illustrate the coupling of transcription and splicing processes, namely the recruitment and the kinetic models. In the recruitment model, a change in promoter architecture results in the recruitment of splicing factors to the transcription machinery that in turn impact the splicing of the nascent RNA. In the kinetic model, the change in promoter architecture affects the elongation rate of the RNAPII, such that there is more or less time for splice sites or other splicing signals flanking the alternative exons to be recognized by trans-acting factors. Thus, this model predicts that elongation rate modulates competition between splice sites and cis-regulatory elements. Such as, if there is a cassette exon flanked by weak upstream 3′-ss and a strong downstream 3′-ss, a lower transcription rate will favor the usage of the upstream site and, consequently, the inclusion of the cassette exon. On the other hand, acceleration of transcription will favor the usage of the downstream site, resulting in exon skipping (Fig. 2.5a). In addition, a lower transcription rate can also favor exon skipping if, for example, an ISE is displayed upstream of a cassette exon. And as mentioned before, the elongation rate can also influence RNA folding events, which contributes to the AS outcome (Braunschweig et al. 2013).

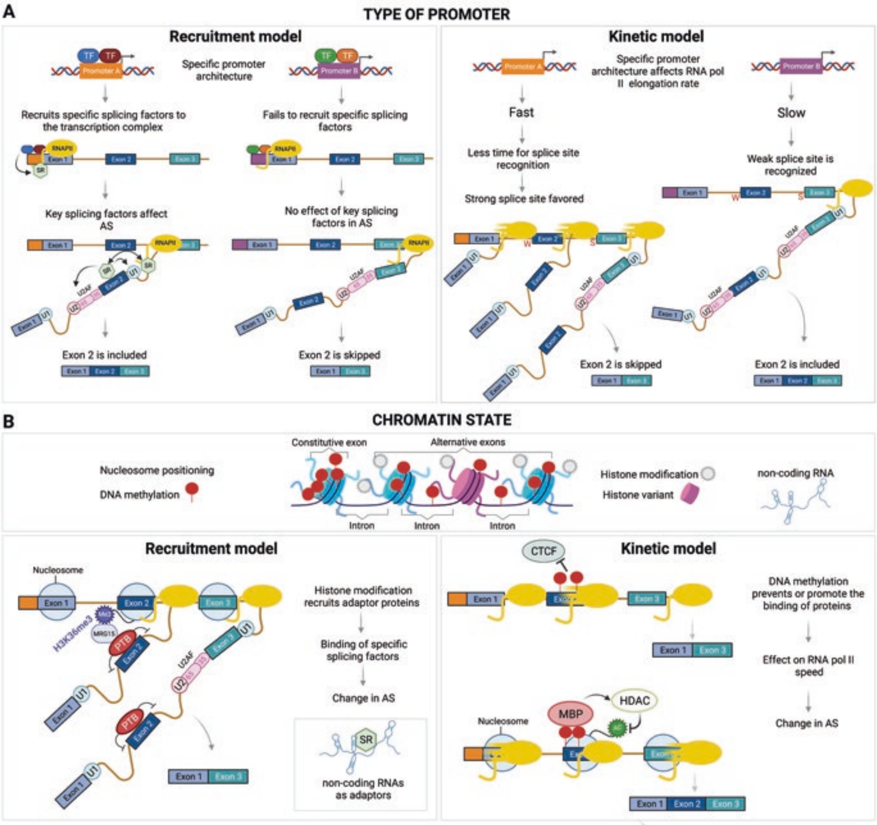

A
TYPE OF PROMOTER
Recruitment model
Specific promoter architecture
Recruits specific splicing factors to the transcription complex
Fails to recruit specific splicing factors
Key splicing factors affect AS
No effect of key splicing factors in AS
Exon 2 is included
Exon 2 is skipped
Kinetic model
Specific promoter architecture affects RNA pol II elongation rate
Fast
Less time for splice site recognition
Strong splice site favored
Slow
Weak splice site is recognized
Exon 2 is skipped
Exon 2 is included
B
CHROMATIN STATE
Constitutive exon
Alternative exons
Nucleosome positioning
DNA methylation
Histone modification
Histone variant
non-coding RNA
Intron
Recruitment model
Nucleosome
H3K36me3
MRG15
PTB
Histone modification recruits adaptor proteins
Binding of specific splicing factors
Change in AS
non-coding RNAs as adaptors
Kinetic model
CTCF
DNA methylation prevents or promote the binding of proteins
Effect on RNA pol II speed
Change in AS
MBP
HDAC

Fig. 2.5 *Promoter and chromatin features affect AS through recruitment and kinetic models.* (**a**) Recruitment model (left): the type of promoter will determine the transcription factors (TFs) that will be recruited to the transcription initiation complex. Promoter A recruits TFs that interact with splicing factors (for example, SR proteins) that will enhance splicing of the alternative exon. The TFs on Promoter B do not interact with splicing factors, affecting the splicing outcome differently from Promoter A. Kinetic model (right): the type of promoter will determine RNAPII elongation rate. Promoter A determines a high elongation rate of the RNAPII, allowing less time to the CTD-associated splicing factors to recognize weak (W) splice sites. Higher RNAPII elongation rate favors recognition of strong (S) splice site. Promoter B determines a slower RNAPII elongation rate, allowing for weak splice site recognition and exon inclusion. (**b**) The chromatin architecture is dictated by many epigenetic layers, such as nucleosome positioning, DNA methylation, histone modification, histone variants, and also non-coding RNAs. All these aspects of chromatin influence alternative splicing. Recruitment model (left): histone modification (such as H3K36me3) recruits an adaptor protein (MRG15) which in turn recruits a splicing factor (PTB, an hnRNP protein with repressor activity) that will affect splicing of the alternative exon. Non-coding RNAs, such as lncRNAs, can act as a scaffold to recruit or sequestrate specific splicing factors. Kinetic model (right): chromatin features affecting the RNAPII elongation rate. For example, DNA methylation prevents CTCF binding, which acts as a roadblock for RNAPII. Without the binding of CTCF, elongation is accelerated and disfavors alternative exon inclusion. DNA methylation on alternative exons can also recruit methyl-binding proteins (MBP) such as MeCP2. MBP recruits histone deacetylase complex (HDAC), leading to less permissive chromatin and slower RNAPII elongation, favoring alternative exon inclusion. (Made in ©BioRender – biorender.com)

2.3.3 Epigenetic Control of Alternative Splicing

Transcription rate is influenced by the binding of transcription factors and co-factors to promoters and regulatory sequences on the DNA, which in turn is shaped by the chromatin structure and the intricate interplay of epigenetic modifications.

The various layers of epigenetic control – DNA methylation, nucleosome positioning, histone modifications, histone variants, chromatin remodeling factors, and non-coding RNAs – are involved in the regulation of AS.

The 5′ cytosine methylation, deposited by DNA methyltransferases (DNMTs) on CpG dinucleotides, corresponds to an essential epigenetic modification on the DNA that influences gene expression patterns across the genome. While deposition of the 5-methylcytosine (5mC) at promoter regions exerts an inhibitory effect on gene expression, its presence on the genes bodies positively affects transcription of the related genes, besides preventing spurious transcription initiation from cryptic internal promoters. 5mC are enriched at exons and especially at splice sites when compared to flanking introns. Moreover, DNA methylation is less abundant in alternatively spliced exons than in constitutive exons (Lev Maor et al. 2015). DNA methylation can either enhance or silence exon recognition. There are three different mechanisms by which DNA methylation regulates AS. DNA methylation can prevent the DNA binding protein, CTCF, from interacting to its binding site, counteracting its function as a roadblock for RNA pol II, culminating in increased elongation rate and exon skipping. Another mechanism involves the binding of methyl-binding proteins, such as MePC2. Binding of MePC2 to methylated DNA triggers recruitment of histone deacetylase complex (HDAC), local hypoacetylation, and consequent RNA pol II pause, favoring exon inclusion (Lev Maor et al. 2015). In a third mechanism, DNA methylation on alternative exons induces the H3K9me3 histone modification, which in turn anchors Heterochromatin Protein 1 (HP1) isoforms HP1α and HP1β. HP1α and HP1β act as adaptor proteins for the recruitment of splicing factors. The presence or absence of HP1 and its associated splicing factors determine whether a cassette exon is included or excluded from the transcript (Yearim et al. 2015).

Nucleosomes are the basic units of chromatin, comprised of an octamer of histones. The DNA wrapped around the nucleosome is approximately 147 nt in length, which is also the average size of exons in mammals. Nucleosomes are enriched in exons, and because of that, they are determinants in exon definition. Moreover, exons that lay between long introns present a higher nucleosome positioning than exons separated by small introns. These facts suggest that nucleosomes may act in both protecting and defining exons (Luco et al. 2011). Alternatively, included exons flanked by weak splice sites present higher nucleosome density than excluded exons, pointing to nucleosome function in AS. Nucleosomes influence AS by acting as a barrier that controls RNA pol II density at exons, lowering transcription rate and favoring the inclusion of cassette exons (Saldi et al. 2016).

Post-translational modification of histones is an important determinant of chromatin. Histone modification such as H3 lysine 36 trimethylation (H3K36me3) and H3 lysine 9 trimethylation (H3K9me3) also influence alternative splicing (Luco et al. 2011). H3K36me3 is highly abundant in actively transcribed genes and is deposited by the methyltransferase SETD2. Dysfunction in this enzyme leads to changes in alternative splicing events. H3K36me3 mark influences splicing by anchoring the chromatin-binding protein, MRG15, which recruits the polypyrimidine binding protein (PTB), a splicing repressor. Therefore, the levels of H3K36me3 will determine the ultimate effect of PTB, favoring exon inclusion or skipping (Luco et al. 2010). Likewise, the H3K9me3 mark also promotes exon skipping, in this case, by recruiting the Heterochromatin protein 1 (HP1). HP1 isoforms HP1α and HP1β act as adaptors to recruit the protein SRSF3. The SRSF3 protein is a member of the SR family that functions as a splicing silencer hindering the inclusion of cassette exons.

As previously mentioned, the H3K9me3 mark recruits HP1, but different from HP1α and HP1β, the isoform HP1γ has a different mode of action. HP1γ binds to the H3K9me3 marks in the gene's coding region and simultaneously associates with the pre-mRNA. Interaction of HP1γ with the nascent pre-mRNA slows RNAPII and consequently the elongation rate. Decrease in the RNAPII elongation rate allows time to recruit splicing factors and cassette exon inclusion (Saint-André et al. 2011).

Additionally, histone variants play a role in AS. The variants H3.3A, H3.3B, H2a.V, and the H3-histone chaperone Asf1 play a role in the processing of histone RNAs itself. It has been described in human lung fibroblasts that deposition of the linker histone variant H1.5 at the splicing sites of short exons constitutes a mark responsible for considerable stalling of RNAPII. Again, decreased RNAPII elongation rate facilitates the inclusion of alternative exons (Glaich et al. 2019). Similarly, BRM – the ATPase subunit of the switch/sucrose non-fermenting, SWI/SNF – a chromatin remodeling factor, facilitates the inclusion of alternative exons by interacting with RNAPII and inducing its pause (Batsché et al. 2006).

In conclusion, the dynamics of the chromatin imposed by DNA methylation, nucleosome positioning, and histone modifications control alternative splicing through the mechanisms portrayed by the recruitment and the kinetic models described earlier (Braunschweig et al. 2013) (Fig. 2.5b).

Ultimately, non-coding RNAs are epigenetic regulators that can affect the chromatin structure and also alternative splicing. Long non-coding RNAs (lncRNAs), such as Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1), taurine upregulated gene 1 (TUG1), and Gomafu, act as a scaffold for the binding of splicing factors, affecting their function in splicing (Ramanouskaya and Grinev 2017). For example, Malat1 associates with SR proteins such as SRSF1, SRSF2, and SRSF3, and its deletion changes the alternative splicing pattern of genes related to tumorigenesis (Zhang et al. 2020).

2.3.4 Coupling Cell Signaling to Alternative Splicing

Alternative splicing also responds to the constant changes in the cell's physiological or pathological conditions (Kornblihtt et al. 2013). In a seminal work, König and colleagues described the mechanism by which splicing can be coupled with signal transduction. The authors found signal-responsive elements within the exonic v5 sequence of the *CD44* pre-mRNA. These ESS and ESE elements modulate alternative splicing in a cell type-specific and inducible manner in response to Ras signaling pathway (König et al. 1998). The mechanism involved Ras-Raf-MEK-ERK activation and SAM68 phosphorylation by ERK. Phosphorylated SAM68 bound to the ESS element resulting in the inclusion of the v5 exon through mechanisms that include impairing the splicing repressor hnRNP A1 from binding to the ESS element, recruiting proteins that promote spliceosome assembly, interacting with chromatin remodeling complex, and decreasing RNAPII elongation rate, and thus favoring the use of weak splice sites (Frisone et al. 2015; Lynch 2007). Together with the protein interactions with the ESE element that enhance splicing, exon v5 is included in the mature mRNA in response to an extracellular signal (Shin and Manley 2004). The TGF-β signaling also controls alternative splicing of the *CD44* pre-mRNA, leading to the expression of the cancer aggressiveness-related isoform CD44v6 (Tripathi et al. 2016).

There are other examples of signaling pathways controlling alternative splicing. The mechanism applied in the majority involves regulation of the SR protein activity through phosphorylation and dephosphorylation events. Such as, Fas receptor activation includes the activity of the phosphatase PP1, which dephosphorylates SR proteins. Altered phosphorylation of the SR splicing factors results in a switch of BCL-X and Caspase 9 transcripts from anti- to pro-apoptotic isoforms. Likewise, phosphoinositide signaling leads to dephosphorylation of the SR factors SRSF10. Dephosphorylated SRSF10 interacts with U1 snRNPs interfering with its 5′-ss recognition and impairing splicing. Also, AKT pathway mediates alternative splicing in response to epidermal growth factor (EGF) signaling through phosphorylation of SR proteins (Kornblihtt et al. 2013).

2.4 Missplicing and Disease

Aberrant splicing causes diseases. According to the Human Gene Mutation Database (HGMD), mutations affecting splicing account for one-third of all disease-causing mutations. The mutations trigger aberrant splicing by mechanisms that involve either disruption of splicing signals or cis-acting regulatory elements on the RNA or interference with the function of the trans-acting factors that act on the RNA.

Mutations in regulatory sequences that affect alternative splicing are a widespread cause of human hereditary disease and cancer. These cis-acting mutations disrupt the splicing code in different ways. It can affect splice sites (5′- and 3′-ss), Py-tract, or BPS or create cryptic splicing signals. It can alter sequences that overlap with the secondary structure of the RNA, hampering its formation or creating folding that is not usually there. Moreover, cis-acting mutations can result in the loss or gain of function of splicing enhancers (ESE, ISE) or silencers (ESS, ISS). The consequence of these alterations is the aberrant splicing of the involved genes due to exon skipping, intron retention, activation of cryptic sites, and the altered ratio of skipping/inclusion of cassette exons.

The following selected examples illustrate disease-associated splicing alterations caused by the different types of *cis*-acting mutations. (i) Familial dysautonomia (FD) is a recessive genetic disorder characterized by a point mutation in the vicinity of the 5′-ss on intron 20 of the *IKBKAP* gene. The altered splice site impairs the recognition and base-pairing of the U1 snRNP at the 5′-ss resulting in exon skipping. (ii) Beta-thalassemia is an inherited disorder characterized by reduced expression of the hemoglobin beta chain and severe anemia. One form results from a point mutation present in the intron 1, which generates an alternative 3′-ss in the *HBB* gene. This cryptic site is preferentially used and results in an aberrant isoform. (iii) The frontotemporal dementia and Parkinsonism linked to Chr.17 (FTDP-17) is a neurodegenerative disorder that can be caused by a point mutation in the exon 10 of the *MAPT* gene. This mutation affects an ESS leading to increased inclusion of the exon. (iv) Familial partial lipodystrophy type 2 (FPLD2) is a rare metabolic condition characterized by point mutations in 5′-ss of the *LMNA* gene, resulting in intron retention and consequent regulation of its transcripts expression through NMD (Daguenet et al. 2015; Scotti and Swanson 2016).

Similar to cis-acting mutations, genetic variations that naturally occur in the population, such as single nucleotide polymorphisms (SNPs), also affect the efficiency of alternative splicing. An example of an allele-dependent expression of alternative isoform is observed for the major histocompatibility complex, class II, DQb 1 (HLA-DQB1) gene. The ratio of *DQB1* exon 4 inclusion in the final transcript is determined by differential recognition of the upstream 3′-ss during the early steps of spliceosome assembly. The differential recognition of the 3′-ss results from differences in the RNA sequence due to the SNPs mapped to this region, directly affecting the BPS and Py-tract (Královičová et al. 2004).

In addition to mutations affecting splicing signals on the pre-mRNA, mutations in genes coding for *trans*-acting factors also cause disease. Unlike the *cis*-acting mutations that only affect the compromised gene, trans-acting mutations alter the function of proteins implicated in the splicing machinery and thus convey a pleiotropic effect on large sets of genes.

Disease-associated trans-acting mutations affect genes involved in UsnRNP biogenesis and assembly or formation of UsnRNP aggregates, spliceosome assembly (core spliceosome mutations), and splicing regulation (SR, hnRNP, and RNA binding proteins).

Spinal muscular atrophy (SMA), Clericuzio-type poikiloderma with neutropenia (PN), Retinitis pigmentosa (RP), and Alzheimer’s disease are examples of disorders associated with mutations that interfere with the function of proteins involved in UsnRNP biogenesis. Prader-Willi syndrome and RP are examples of disorders associated with mutations that interfere with core spliceosome- and splicing factor protein. At the same time, amyotrophic lateral sclerosis (ALS), autism disorder, and Huntington's disease have been associated with deregulated expression of splicing factors and RNA-binding proteins.

The following selected examples illustrate disease-associated trans-acting mutations involved in each of the mechanisms described. (i) Clericuzio-type poikiloderma with neutropenia (PN) rare autosomal recessive disease associated with mutations in the *C16orf57* gene. This gene encodes a protein involved in the correct processing of the U6 snRNA. Cells from patients carrying these mutations have higher levels of U6 snRNA degradation. (ii) RP are inherited degenerative disorders of the retina associated with mutations in multiple genes, among them the pre-mRNA processing factors *PRPF31*, *PRPF8*, *PRPF6*, *PRPF3*, and the RNA helicase *SNRNP200/BRR2*. These genes code for components of the spliceosome complex involved in complex rearrangement and catalysis. (iii) ALS is a common motor/neurodegenerative disease caused by mutations in various genes, such as the ones coding for the RNA-binding protein FUS and TDP-43. Their altered function affects snRNA abundance and, consequently, pre-mRNA splicing (Daguenet et al. 2015).

2.5 Splicing and Cancer

In the context of cancer, mutations in the cis-acting regulatory sequences on the RNA often generate aberrant tumor-associated isoforms that contribute to some aspects of tumorigenesis. Additionally, cis-acting mutations can contribute to activate oncogenic isoforms or inactivate tumor suppressor transcripts. On a broader scale, somatic mutations in genes encoding components of the splicing machinery are also frequently observed and are related to global splicing abnormalities of cancer transcriptomes (Dvinge et al. 2016).

Recurrent somatic mutations in core spliceosome and splicing factor coding genes were first discovered in hematological malignancies like myelodysplastic syndromes (MDS), acute myeloid leukemia (AML), and chronic lymphoblastic leukemia (CLL) (Graubert et al. 2012; Papaemmanuil et al. 2011; Quesada et al. 2012; Wang et al. 2011; Yoshida et al. 2011), and later identified with high frequency in a variety of solid tumors, such as uveal melanoma (Harbour et al. 2013), lung adenocarcinoma (Imielinski et al. 2012), breast (Maguire et al. 2015; Stephens et al. 2012), and pancreatic cancer (Biankin et al. 2012). These findings were the first direct genetic link between dysfunction of splicing machinery and cancer, and defects in this machinery have been proposed as leukemogenic pathways (Maciejewski and Padgett 2012).

Interestingly, these mutations are heterozygous and mutually exclusive, indicating that cells may tolerate only partial deviation from normal splicing. In fact, cells carrying splicing factor mutations are sensitive to genetic or pharmacological perturbation of splicing (Fei et al. 2016; Obeng et al. 2016; Seiler et al. 2018; Shirai et al. 2017; Zhou et al. 2015).

The most frequently reported splicing mutations in cancer occur in four genes, namely *SF3B1*, *SRSF2*, *U2AF1*, and *ZRSR2*. At least for *SF3B1*, *U2AF1*, and *SRSF2*, mutations affect highly restricted residues within the proteins, suggesting again (or alteration) of function phenotype, whereas *ZRSR2* mutations are widespread throughout the protein and follow a loss of function pattern. These mutations affect splicing by interfering with 3′-ss recognition mediated by *SF3B1; U2AF1* on U2-type introns or by *ZRSR2* on U12-type introns as well as with exon recognition mediated by SRSF2 (Dvinge et al. 2016).

Among the less frequently mutated splicing factors genes are *SF1*, *U2AF2*, *SF3A1*, *PRPF40B* (Yoshida et al. 2011), *PRPF8*, *LUC7L2*, *HCF1*, *SAP130*, *SRSF6*, *SON*, and *U2AF26* (Makishima et al. 2012). Together with *SF3B1*, *SRSF2*, *U2AF1*, and *ZRSR2*, these genes encode multiple components or associated factors from the spliceosome complexes E/A (Fig. 2.6). Although less frequently mutated, these

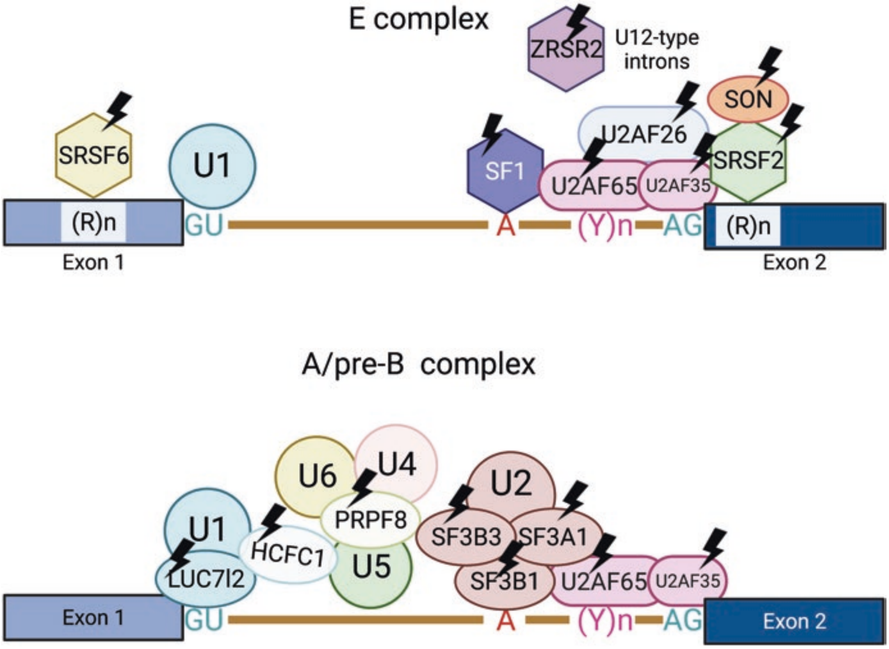


Fig. 2.6 *Core spliceosome and splicing factor proteins affected by somatic mutation in cancer*. A lightning symbol represents the affected components. E complex: U2AF35 binds to AG dinucleotide on the 3′-ss of the intron, while U2AF65 binds to the Py-tract, represented by (Y)n. SF1 binds to the BPS, represented by an A. U2AF26 interacts with U2AF to perform essential functions in splicing. ZRSR2 acts on 3′-ss recognition of U12-type introns. Arginine/serine-rich splicing factors SRSF2 and SRSF6 bind to polypurine sequences ((R)n) in the exon. SRSF2 interacts with U2AF65. SON, a recently discovered spliceosomal gene, interacts with SRSF2 and mediates constitutive splicing of weak splice sites. A/pre-B complex: SF1 is replaced by U2 snRNP along with its components SF3A1, SF3B1, and SF3B3. LUC7L2 is associated with the U1 snRNP on 5′-ss. PRPF8 plays an essential role in the interaction among U4/U6/U5 snRNPs, while HCFC1 contributes to the U1/U5 interaction. (Made in ©BioRender – biorender.com)

genes participate in the same molecular pathway as the frequently mutated genes, indicating that the impairment of the pathway rather than the individual molecule is important for carcinogenesis.

In addition to mutations affecting *trans*-acting factors, the differential expression of splicing regulatory factors and altered post-translational modification of these proteins are strongly associated with splicing abnormalities and transformation. Mounting evidence suggests that these factors can act as both oncoproteins and tumor suppressors. Among the growing list are the SR proteins, hnRNPs, and other splicing factors such as SRSF1, SRSF3, SRSF6, SRSF10, hnRNP A1, hnRNP A2, hnRNP A2/B1, hnRNP H, hnRNP K, hnRNP A2 hnRNP M, PRPF6, PTB QKI, RBFOX2, RBM4, RBM5, RBM6, and RBM10 (for review, see Dvinge et al. 2016; Grosso et al. 2008).

The first mechanistic evidence that deregulated splicing factor expression resulted in the malignant transformation was demonstrated for the SR factor, SRSF1. SRSF1 is upregulated in several tumors, and this is sufficient to affect the alternative splicing of the BIN1 tumor suppressor and the MNK2 and S6K1 kinases. The resulting isoform of BIN1 has no tumor suppressor activity, whereas those of MNK2 and S6K1 have shown oncogenic properties (Karni et al. 2007).

There is an overlap between the splicing factors with altered expression in cancer, which are also mutated in hematological malignancies, such as *U2AF1*, *SRSF2*, *SFRS6*, and *SF1* (Grosso et al. 2008), indicating that disturbing the function of these proteins at any level might contribute to disease.

A major challenge in research today is to associate the mutations and aberrant expression/activity of the splicing factors with specific downstream splicing changes.

References

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, Kim T, Misquitta-Ali CM, Wilson M, Kim P, Odom DT, Frey B, Blencow B (2012) The evolutionary landscape of alternative splicing in vertebrate species. Science 338(80):1587–1593

Batsché E, Yaniv M, Muchardt C (2006) The human SWI/SNF subunit Brm is a regulator of alternative splicing. Nat Struct Mol Biol 13:22–29

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414

Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, Patch AM, Wu J et al (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491:399–405

Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ (2013) Dynamic integration of splicing within gene regulatory pathways. Cell 152:1252–1269

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ (2014) Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res 24:1774–1786

Cáceres JF, Stamm S, Helfman DM, Krainer AR (1994) Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. Science 265(80):1706–1709

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298

Cordin O, Beggs JD (2013) RNA helicases in splicing. RNA Biol 10:83–95

Daguenet E, Dujardin G, Valcárcel J (2015) The pathogenicity of splicing defects: mechanistic insights into pre- mRNA processing inform novel therapeutic approaches. EMBO Rep 16:1640–1655

David CJ, Manley JL (2011) The RNA polymerase II C-terminal domain-a new role in spliceosome assembly. Transcription 2:227–231

De Conti L, Baralle M, Buratti E (2013) Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip Rev RNA 4:49–60

Dvinge H, Kim E, Abdel-Wahab O, Bradley RK (2016) RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer 16:413–430

Erkelenz S, Mueller WF, Evans MS, Busch A, Schöneweis K, Hertel KJ, Schaal H (2013) Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. RNA 19:96–102

Fei DL, Motowski H, Chatrikhi R, Prasad S, Yu J, Gao S, Kielkopf CL, Bradley RK, Varmus H (2016) Wild-type U2AF1 antagonizes the splicing program characteristic of U2AF1-mutant tumors and is required for cell survival. PLoS Genet 12:1–26

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc Natl Acad Sci U S A 102:16176–16181

Frisone P, Pradella D, Di Matteo A, Belloni E, Ghigna C, Paronetto MP (2015) SAM68: signal transduction and RNA metabolism in human cancer. Biomed Res Int

Geuens T, Bouhy D, Timmerman V (2016) The hnRNP family: insights into their role in health and disease. Hum Genet 135:851–867

Glaich O, Leader Y, Lev Maor G, Ast G (2019) Histone H1.5 binds over splice sites in chromatin and regulates alternative splicing. Nucleic Acids Res 47:6145–6159

Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE et al (2012) Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nat Genet 44:53–57

Grosso AR, Martins S, Carmo-Fonseca M (2008) The emerging role of splicing factors in cancer. EMBO Rep 9:1087–1093

Hall SL, Padgett RA (1994) Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J Mol Biol 239:357–365

Harbour JW, Roberson EDO, Anbunathan H, Onken MD, Worley LA, Bowcock AM (2013) Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. Nat Genet 45:133–135

Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A et al (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 150:1107–1120

Karni R, De Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR (2007) The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol 14:185–193

König H, Ponta H, Herrlich P (1998) Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator. EMBO J 17:2904–2913

Kornblihtt AR (2005) Promoter usage and alternative splicing. Curr Opin Cell Biol 17:262–268

Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol 14:153–165

Královičová J, Houngninou-Molango S, Krämer A, Vořechovský I (2004) Branch site haplotypes that control alternative splicing. Hum Mol Genet 13:3189–3202

Lev Maor G, Yearim A, Ast G (2015) The alternative role of DNA methylation in splicing regulation. Trends Genet 31:274–280

Li X, Liu S, Zhang L, Issaian A, Hill RC, Espinosa S, Shi S, Cui Y, Kappel K, Das R et al (2019) A unified mechanism for intron and exon definition and back-splicing. Nature 573:375–380

Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 12:1998–2012

Long JC, Caceres JF (2009) The SR protein family of splicing factors: master regulators of gene expression. Biochem J 417:15–27

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T (2010) Regulation of alternative splicing by histone modifications. Science 327(80):996–1000

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T (2011) Epigenetics in alternative pre-mRNA splicing. Cell 144:16–26

Lynch KW (2007) Regulation of alternative splicing by signal transduction pathways. Adv Exp Med Biol 623:161–174

Maciejewski JP, Padgett RA (2012) Defects in spliceosomal machinery: a new pathway of leukaemogenesis. Br J Haematol 158:165–173

Maguire SL, Leonidou A, Wai P, Marchiò C, Ng CKY, Sapino A, Salomon AV, Reis-Filho JS, Weigelt B, Natrajan RC (2015) SF3B1 mutations constitute a novel therapeutic target in breast cancer. J Pathol 235:571–580

Makishima H, Visconte V, Sakaguchi H, Jankowska AM, Kar SA, Jerez A, Przychodzen B, Bupathi M, Guinta K, Afable MG et al (2012) Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. Blood 119:3203–3210

Manley JL, Krainer AR (2010) A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev 24:1073–1074

Martinez-Contreras R, Cloutier P, Shkreta L, Fisette J, Revil T, Chabot B (2007) hnRNP proteins and splicing control. Adv Exp Med Biol 623:123–147

Matera AG, Wang Z (2014) A day in the life of the spliceosome. Nat Rev Mol Cell Biol 15:108–121

McManus CJ, Graveley BR (2011) RNA structure and the mechanisms of alternative splicing. Curr Opin Genet Dev 21:373–379

Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science 338(80):1593–1599

Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457–463

Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, Schneider RK, Lord AM, Wang L, Gambe RG et al (2016) Physiologic expression of Sf3b1K700E causes impaired erythropoiesis, aberrant splicing, and sensitivity to therapeutic spliceosome modulation. Cancer Cell 30:404–417

Padgett RA, Grabowski PJ, Konarska MM, Seiler S, Sharp PA (1986) Splicing of messenger RNA precursors. Annu Rev Biochem 55:1119–1150

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415

Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C et al (2011) Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med 365:1384–1395

Patel AA, Steitz JA (2003) Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 4:960–970

Pervouchine DD, Khrameeva EE, Pichugina MY, Nikolaienko OV, Gelfand MS, Rubtsov PM, Mironov AA (2012) Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. RNA 18:1–15

Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A et al (2012) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet 44:47–52

Ramanouskaya TV, Grinev VV (2017) The determinants of alternative RNA splicing in human cells. Mol Gen Genomics 292:1175–1195

Reed R (1996) Initial splice-site recognition and pairing during pre-mRNA splicing. Curr Opin Genet Dev 6:215–220

Saint-André V, Batsché E, Rachez C, Muchardt C (2011) Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat Struct Mol Biol 18:337–344

Saldi T, Cortazar MA, Sheridan RM, Bentley DL (2016) Coupling of RNA polymerase II transcription elongation with pre-mRNA splicing. J Mol Biol 428:2623–2635

Scotti MM, Swanson MS (2016) RNA mis-splicing in disease. Nat Rev Genet 17:19–32

Seiler M, Yoshimi A, Darman R, Chan B, Keaney G, Thomas M, Agrawal AA, Caleb B, Csibi A, Sean E et al (2018) H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. Nat Med 24:497–504

Sharp PA (1994) Split genes and RNA splicing (nobel lecture). Angew Chem Int Ed Eng 33:1229–1240

Sharp PA (2005) The discovery of split genes and RNA splicing. Trends Biochem Sci 30:279–281

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R (2006) Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res 34:3955–3967

Shin C, Manley JL (2004) Cell signalling and the control of pre-mRNA splicing. Nat Rev Mol Cell Biol 5:727–738

Shirai CL, White BS, Tripathi M, Tapia R, Ley JN, Ndonwi M, Kim S, Shao J, Carver A, Saez B et al (2017) Mutant U2AF1-expressing cells are sensitive to pharmacological modulation of the spliceosome. Nat Commun 8

Smith CWJ, Valcárcel J (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci 25:381–388

Staley JP, Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell 92:315–326

Stephens PJ, Tarpey PS, Davies H, Van Loo P, Wedge DC, Nik-zainal S, Martin S, Varela I, Bignell GR, Yates LR et al (2012) Stephens-2012-somatic landscape breast ca. 486:400–404

Tripathi V, Sixt KM, Gao S, Xu X, Huang J, Weigert R, Zhou M, Zhang YE (2016) Direct regulation of alternative splicing by SMAD3 through PCBP1 is essential to the tumor-promoting role of TGF-β. Mol Cell 64:549–564

Wahl MC, Will CL, Lührmann R (2009) The spliceosome: design principles of a dynamic RNP machine. Cell 136:701–718

Wan R, Bai R, Shi Y (2019) Molecular choreography of pre-mRNA splicing by the spliceosome. Curr Opin Struct Biol 59:124–133

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476

Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L et al (2011) SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med 365:2497–2506

Warf MB, Berglund JA (2010) Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem Sci 35:169–178

Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, Nissim-Rafinia M, Cohen AHS, Rippe K, Meshorer E et al (2015) HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. Cell Rep 10:1122–1134

Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M et al (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. Nature 478:64–69

Zhang C, Li WH, Krainer AR, Zhang MQ (2008) RNA landscape of evolution for optimal exon and intron discrimination. Proc Natl Acad Sci U S A 105:5797–5802

Zhang J, Zhang YZ, Jiang J, Duan CG (2020) The crosstalk between epigenetic mechanisms and alternative RNA processing regulation. Front Genet 11:1–10

Zhou Q, Derti A, Ruddy D, Rakiec D, Kao I, Lira M, Gibaja V, Chan HM, Yang Y, Min J et al (2015) A chemical genetics approach for the functional assessment of novel cancer genes. Cancer Res 75:1949–1958

Chapter 3
Transcriptome Analysis Using RNA-seq and scRNA-seq

Waldeyr Mendes Cordeiro Silva, Fabián Andrés Hurtado, Kelly Simi, Pedro Henrique Aragão Barros, Dimitri Sokolowskei, Ildinete Silva-Pereira, Maria Emilia Walter, and Marcelo Brigido

3.1 High-Throughput Sequencing Techniques

Since Sanger's technology in the 1970s, DNA sequencing has been continuously improved regarding both throughput and low cost. Next-generation sequencing (NGS), also called high-throughput or deep sequencing, constitutes a new breakthrough in increasing research power, a revolutionary advancement in molecular biology knowledge. An increasing number of biological questions may be addressed by NGS technologies, which provide a much larger comprehensive survey compared to the Sanger method, and under a system biology perspective. Transcriptomics has been particularly benefited by the use of these new technologies, also called RNA-seq, allowing a complete characterization of the whole transcriptome at both gene (Kvam et al. 2012) and exon (Anders et al. 2012) levels,

W. M. C. Silva
Laboratório de Biologia Molecular, CEL/IB, Universidade de Brasília, Brasília, DF, Brazil

Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Formosa, GO, Brazil

F. A. Hurtado · P. H. A. Barros · D. Sokolowskei
Laboratório de Biologia Molecular, CEL/IB, Universidade de Brasília, Brasília, DF, Brazil

Graduation program in Molecular Pathology/FM/UnB, Universidade de Brasília, Brasília, DF, Brazil

K. Simi
UniCEUB, Centro Universitario de Brasília, Brasília, DF, Brazil

I. Silva-Pereira · M. Brigido (✉)
Laboratório de Biologia Molecular, CEL/IB, Universidade de Brasília, Brasília, DF, Brazil
e-mail: brigido@unb.br

M. E. Walter
Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília, Brasília, DF, Brazil

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_3

and with an additional ability to identify rare transcripts, new genes, novel splicing junctions, and gene fusions (Wang et al. 2009; Katz et al. 2010; Van Verk et al. 2013). More recently, single-cell sequencing had become a feasible task allowing a deeper and systemic view of individual cell's transcriptomes.

This chapter first addresses a brief overview of sequencing techniques and the most common next-generation platforms and computational methods for RNA-seq data analysis. Then, we present two case studies to assess the capabilities of RNA-seq in addressing important biological issues.

3.1.1 Sanger's Sequencing Technology

In 1977, Frederick Sanger and colleagues (1977) developed the DNA sequencing method, which in 2001 allowed the first human genome draft (Lander et al. 2001). This method, called dideoxy chain-termination or simply the Sanger method, is based on special nucleotide molecules (called ddTNP), lacking a 3′-OH at the deoxyribose, which blocks the DNA elongation. These special nucleotides are mixed in lower concentrations to the regular nucleotides and used as reagents for DNA polymerase reaction. Therefore, with the polymer synthesis stopped by the ddNTP's inclusion, the last nucleotide can be determined. Each of the four ddNTPs was added separately in four different reactions. In the beginning, one of the regular nucleotides, most commonly dATP or dCTP, was radioactively labeled (e.g., 32P or 35S) to achieve the radioactive signal. Usually, polyacrylamide gel electrophoresis was used to separate the DNA molecules, which diverged in length by a single nucleotide. Then the gel was dried and exposed to X-ray film.

An important modification of the method was substituting the radioactive label with a fluorescent dye (Smith et al. 1986). Each distinct wavelength produced by the fluorescent dyes linked to dideoxynucleotides corresponds to a different nucleotide, with the four sequencing reactions performed in the same tube. With the Sanger sequencing method's automation, the performance reached up to 96 different reactions running in parallel capillary gel electrophoresis (Marsh et al. 1997), which is considered the first-generation technology. At the top of the technology, 384 samples could be sequenced at once in a single multi-well plate. The Sanger method's main sequencing devices are ABI (Applied Biosystems) and MegaBACE (GE Healthcare Life Sciences).

3.1.2 Next Generation Sequencing

Regulatory mechanisms and gene expression profiles have been widely investigated toward the elucidation of several essential cellular processes. Hybridization-based technology, e.g., microarray, has been beneficial for determining global gene expression. However, the high background levels due to cross-hybridization, a limited range of quantification, and a restricted detection of known genes are bottlenecks for large-scale use of this technique (Shendure 2008). RNA-seq allows a genome-scale transcriptome analysis, including novel genes and splice variants, with a wide range of quantification and reduced sequencing costs (Wang et al. 2009; Soon et al. 2013). These advantages make RNA-seq a better and attractive solution for whole-genome transcriptome analysis of several organisms, even for those with no sequenced reference genomes.

Nowadays, the most commonly used NGS platforms for RNA-seq research are Illumina, PacBio, and Nanopore. These and other novel platforms are rapidly becoming more popular as they profile short and longer reads at a reasonable price per base. The substitution of older NGS technology is fast and pioneer methods, such as pyrosequencing, are nowadays wholly abandoned. A comparison of current NGS technologies is shown in Table 3.1.

The enormous amounts of data generated by NGS create new challenges to the downstream bioinformatics analysis, which has to handle large sequence files while searching for comprehensive and useful biological information, discussed later in this chapter.

3.1.3 Illumina Sequencing

Illumina sequencing uses a reversible dye-terminator technique that adds a single nucleotide to the DNA template in each cycle (Bentley et al. 2008). This system was initially developed in 2007 by Solexa and was subsequently acquired by Illumina, Inc. Illumina is widely used in several transcriptome studies since it reaches the deepest depth among NGS technologies, despite its small sequence size (150–300 bp).

Illumina sequencing is based on sequencing-by-synthesis. Sequencing is performed in a solid slide covered by adaptors complementary to those added to the fragmented DNA sequences (Metzker 2010). This procedure, called bridge PCR, consists of amplifying bent DNA sequences attached by both ends to the solid surface (Fig. 3.1a). By the end of the clonal amplification, clusters of identical DNA sequences (Polonies) will be formed to amplify the fluorescence signals. In each round, one single nucleotide is added to the single-strand template sequences followed by fluorescence detection by a high-sensitivity CCD camera (Fig. 3.1b).

Table 3.1 Comparison of next-generation sequencing technologies

	ABI 3730xl (Sanger)	Illumina	PacBio	Nanopore
Read length (bp)	900	75–300	5000–60000	500–2300000
Cost (US$/Mb)	500	0.01–0.063	0.013–0.933	0.021–2
Output data/run	2,88 Mb	~120 Gb	2–160 Gb	10–300 Gb
Time run (hours)	3	12–44	Up to 4 h	0.017–72

Data from Amarasinghe et al. (2020); Logsdon et al. (2020)

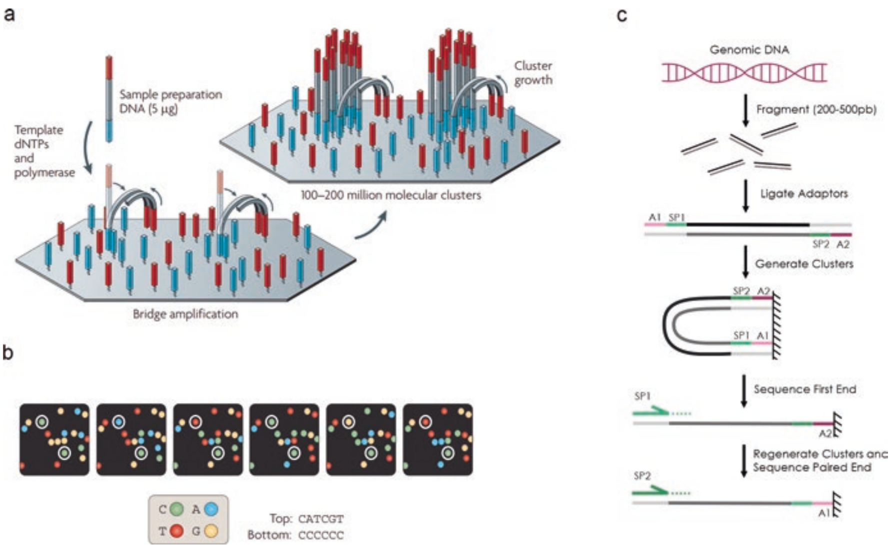


Fig. 3.1 The Illumina sequencing technology. (**a**) Two basic steps encompass an initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template in a solid device with immediately adjacent primers to form clusters; (**b**) In the images, the sequencing data is highlighted from two sequence clusters; (**c**) Paired-end sequencing by which reads are generated from both template strand. "A" block indicates the device-ligation adaptors and "SP," sequencing primers. (Source: Metzker (2010) and http://www.illumina.com/)

As in Sanger's technology, different fluorophore molecules are attached to each nucleotide; however, only one nucleotide is incorporated in each cycle. The fluorescence emission releases the 3′OH of the recently added nucleotide, allowing it to receive new monomers in the subsequent sequencing cycle.

Single-end sequencing, i.e., reads generated from a single-end adaptor, is being replaced by the paired-end sequencing since the accuracy of downstream analysis is greater with a fair price. Paired-end reads are produced from the adaptor priming sites in both template sequence ends, the second adaptor primer being used in a subsequent sequencing run (Fig. 3.1c).

3.1.4 Pacific BioSciences Sequencing

Single-molecule real-time (SMRT) sequencing was devised by Pacific BioSciences (PacBio) in 2009, and it is also called PacBio sequencing (Eid et al. 2009). This platform uses a single DNA polymerase attached to the bottom of a picolitre well – zero-mode waveguides (ZMW) – which replicates a single-molecule template per well to produce a signal for light detection in the smallest volume. In this method, the template is capped by hairpin adapters at both ends of the double-stranded DNA molecule, forming a single-stranded circular DNA (called a SMRTbell). Consequently, the polymerase repeatedly passes over the circular template and sequencing it multiple times, resulting in long read lengths and, thus, providing higher accuracy (Rhoads and Au 2015). The PacBio platform enables simultaneous analysis of millions of wells per chip in a single run, providing long read lengths to up to 60 kb (with average read lengths of 20 kb) (Nakano et al. 2017).

Overall, this technology is considered highly accurate and robust, even as its first sequencers have some drawbacks that narrow down its application. For instance, the limited high-throughput, higher cost, and error rate compared with those of second-next generation sequencing (SGS) technologies (Kanzi et al. 2020; Wang et al. 2020). However, in 2019, PacBio launched the Sequel II System, which asserts improvements in the sequencing to deal with these limitations, generating highly accurate (99.9%) individual long reads up to 25 kb (HiFi reads) and reduces the costs and time of the project, in comparison with its prior versions (Wenger et al. 2019; Logsdon et al. 2020). These HiFi reads are generated by using the circular consensus sequencing (CCS) due to continuous circular sequencing (Wenger et al. 2019; Pereira et al. 2020).

For transcriptomic analysis, the SMRT isoform sequencing (Iso-Seq) from PacBio increased the read length compared to other SGS technologies. This platform achieves full-length transcripts sequencing, improving the analysis in different applications, including gene annotation, isoform identification, fusion transcripts identification, and long non-coding RNA discovery (Weirather et al. 2015; Nattestad et al. 2018; Wang et al. 2019; Zhang et al. 2020a; Hu et al. 2020).

3.1.5 Nanopore MinION Sequencing

The long-read-length sequencer MinION, the first nanopore sequencer device, was announced by Oxford Nanopore Technologies (ONT) in 2012 as a portable, compact, real-time sequencing controlled by a laptop computer device (Deamer et al. 2016). Since then, new nanopore platforms have quickly emerged, such as PromethION, which offers a greater scale of sequencing, and SmidgION, the smallest sequencing platform designed for use with smartphones or other mobile devices.

After library preparation, each strand is attached to adapters. The adaptors bind to a protein motor that guides the sequence to the protein pore, which processes it. Beginning at the 5′-end, the DNA or RNA polymer passes through the pore controlled by the motor protein, which unzips dsDNA and translocates a single strand sequence (Fig. 3.2). The translocated strand modulates the ion current flow through the pore membrane (Ip et al. 2015). The variation of the electrochemical current promoted by each different nucleotide is measured by a sensor and enables identification by different signal patterns. The resultant signals are stored in a FAST5 format file and can be finally used for base-calling, a process in which the nucleotides are predicted from the Raw signals and transferred to a FASTQ file.

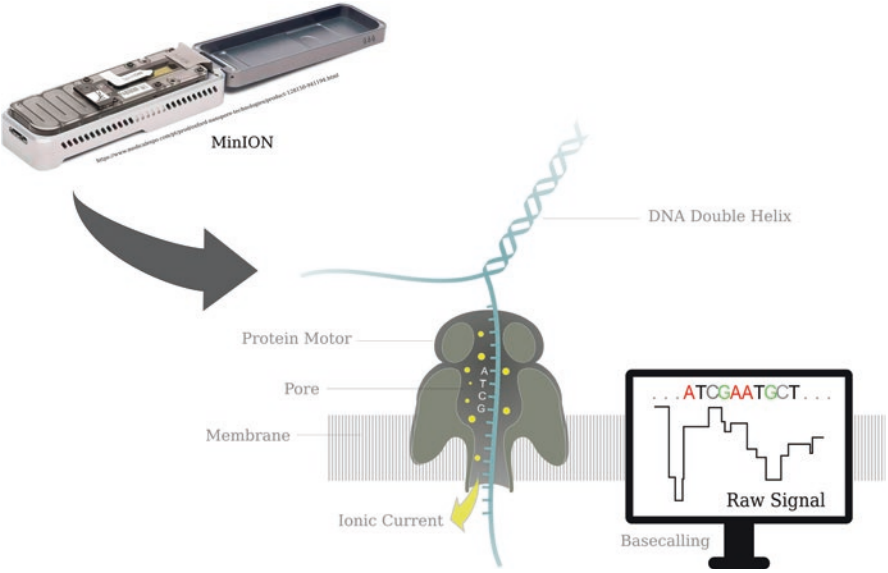


Fig. 3.2 Schematic view of the nanopore sequencer. MinION device process double DNA helix. First, the protein motor unzips DNA passing a single strand through the pore. The movement of the single strand promotes an ionic current flow that is measured and converted to nucleotides data by the base calling analysis

Base-calling can be performed using only information from one strand (1D) or two strands (2D) for consensus, with information from both strands resulting in better base prediction (Lu et al. 2016). Currently use of neural networks in base-calling reached an accuracy interval between 85% and 95% with the detection of signal patterns (Zhang et al. 2020b).

Although sequencing full-length reads allows improvement of isoforms identification and discovery in transcriptome sequencing, it deals with high error rates (Kovaka et al. 2019). To reduce error rates before analysis, nanopore correcting errors can be made by a hybrid error correction strategy. This strategy uses high accuracy short reads to correct long-reads, self-correction methods that rely only on long-reads, or reference-based methods that use a reference genome for error correction (Zhao et al. 2019).

3.2 Bioinformatics Pipelines for Transcriptome Projects

Illumina sequencing is the most used technique in transcriptome studies, since the number of sequenced reads (named raw data) allows to find out virtually the complete set of expressed genes (transcripts). However, longer reads allow a more precise definition of the transcripts. In both cases, the metaphor for reconstructing the transcripts is like mounting a puzzle, where the pieces (the reads) have to be assembled (relative to a reference genome or not) to obtain the picture (transcripts in a transcriptome). After this, different analyses can be performed on these reconstructed transcripts, e.g., quantitative and differential expression. In a transcriptomic project, the tasks of reconstructing transcripts and performing biological analyses are performed by bioinformatics pipelines, discussed next.

3.2.1 Pipelines

A bioinformatics pipeline or workflow is a computational system composed of a sequence of programs sequentially executed. The output data from one software is the input data for the following software (Wercelens et al. 2019). In general, transcriptome bioinformatics pipelines have the following steps, which can be combined according to the raw input data and the objectives of each project:

- Quality control of raw data: This initial step allows visualization, analysis, and filtering (cleaning) the data. Usually, this process takes two sub-steps as follows: clipping and trimming. In the clipping step, adapters (primers) attached to the ends of the sequenced reads (or even the whole read) are removed. In the trimming step, low-quality sequences in the reads are filtered. The filtering guarantees a reliable dataset of quality reads to be used in the following phases of the pipeline.
- Assembly: in the absence of a reference genome or transcriptome, it is necessary to assembly one. For that, overlapping reads (the end of a read is similar to the beginning of another read) are joined in groups of reads (called contig), allowing to construct of one larger sequence (called consensus), which is a predicted (fragment of) transcript. The complete set of transcripts is the predicted transcriptome (Fig. 3.3).
- Mapping: The filtered reads can be aligned to the transcriptome's reference genome to find the actively expressed exons or transcripts. The amount of reads mapping to a single exon/transcript is proportional to its expression.
- Analysis: The whole set of (fragments of) transcripts obtained from the mapping or the assembling step allows to obtain relevant biological information, e.g.

  (a) quantitative analysis: among others, coverage analysis shows the abundance of genes expressed in one RNA-seq sample, more precisely, the number of reads mapped in a certain region of the chromosome.
  (b) differential expression: allows to analyze the differences and variability of gene expression between samples along distinct genomic regions.
  (c) annotation: assigns a biological function to each transcript.

Designing a particular pipeline mainly depends on the transcriptome project's objectives and other information, such as the sequencing platform employed (since the sequencing techniques may cause specific errors in the raw data). It also depends

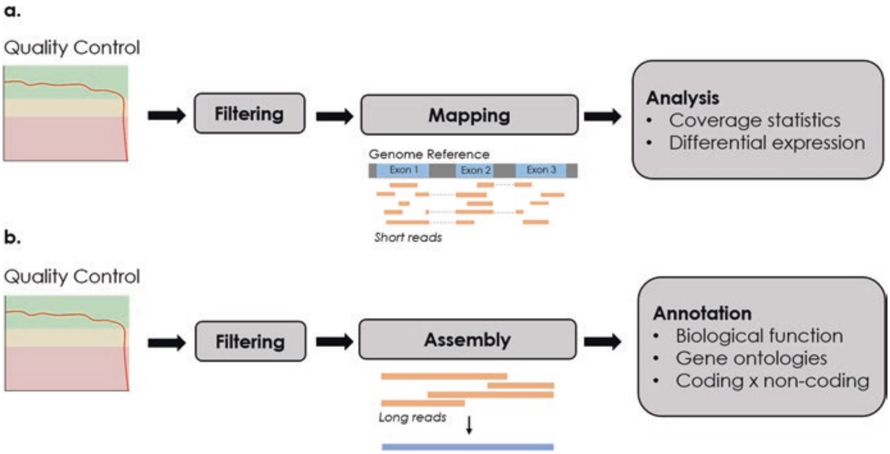


Fig. 3.3 Examples of pipelines for transcriptome analysis: (**a**) Pipelines for short reads, with a well-characterized reference genome, and two types of analyses – coverage statistics and differential expression. (**b**) Pipeline for longer reads, with no reference genome, and annotation (biological function, gene categories, and ontologies)

on the availability of a reference genome or transcriptome in the mapping step and the analysis step's accuracy and biases. Two generic bioinformatics pipelines for transcriptomes are discussed next.

Pipeline 1 The organisms of interest have already been sequenced, preferably with high coverage, well-annotated genes, and other relevant biological characteristics. The reads are usually short (about 150–300 bp), typically produced by Illumina sequencing platforms. This pipeline would be composed of a minimum of three steps (Fig. 3.3a): quality control, mapping, and quantitative analysis.

Pipeline 2 The organism of interest has not been sequenced before. The reads are usually long (up to 40 kb), heavily produced by the PacBio sequencing platform. This pipeline would be composed of a minimum of three steps (Fig. 3.3b): quality control, assembly, and annotation. The assembly phase constructs one consensus sequence for each group of reads presenting similar extremities. This approach heavily depends on sequencing quality, and the multiplatform approach improves the final assembled transcriptome. Finally, the annotation phase assigns biological functions to the consensus sequences.

A bioinformatics pipeline is usually implemented using command lines (e.g., GNU/Linux terminal) mainly because it is a fast, relatively simple, and reliable way to control and manipulate large amounts of datasets. Programming languages such as Shell Script, Python, R, and Perl might also help implement a pipeline and resolve minor tasks by scripting. The pipeline's files/data can be organized in directories or database management systems, relational databases (e.g., MySQL, Oracle), or NoSQL databases (e.g., MongoDB, Neo4J) to store, retrieve, and manage the data.

Most software used in pipelines are free, open-source, publicly available, and some of the most common ones are described next.

Frameworks to manage workflows are also available, such as Snakemake (Köster and Rahmann 2012) and Common Workflow Language (CWL) (https://www.commonwl.org/v1.0). They provide a reliable way to standardize the syntax and semantics for program evoking and create robust and reproducible workflows.

3.2.2 Bioinformatics Software

3.2.2.1 Software for Quality Control

The overall quality of the output sequencing data must be assessed to eliminate bad quality, poorly sequenced, or ambiguous raw data that could negatively impact further analysis. Thus, filtering (or cleaning) strategies capable of clipping and trimming are essential to guarantee the reliability of transcriptomics data and ensure obtaining relevant and trustworthy biological information. The sequenced reads are stored using FASTQ format, gathering the nucleotides sequences of each read and their corresponding quality scores.

Some tools are used to assess and visualize the overall quality of data, such as FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), a popular java-based quality control check program. Other tools to perform filtering steps like FASTX-Toolkit (http://hannonlabcshledu/fastx_toolkit) provide options for performing both clipping and trimming. Other commonly used tools are Cutadapt (Martin 2011) for clipping, PRINSEQ (Schmieder and Edwards 2011), and Trimmomatic (Bolger et al. 2014) for trimming. Fastp (Chen et al. 2018) is an ultra-fast all-in-one quality control, and data-filtering tool that can be an alternative to multiple and insufficiently fast software for quality control. They all present several options, such as minimum size for a read, minimum quality score, and polyadenylation removal.

3.2.2.2 Software for Mapping

The mapping phase's main objective is to find where each filtered short read corresponds in a reference genome/transcriptome (Fig. 3.4).

There are many programs capable of performing the mapping process. In general, these software are computationally intensive (to process and store data), and mapping techniques use indices to accelerate the search procedure and reduce the memory cost associated with finding the location of reads to the reference genome.

Bowtie (Langmead et al. 2009) is a fast short aligner that tolerates a small number of mismatches. Bowtie first concatenates all the reference genome in one single string and performs the Burrows-Wheeler transformation (BWT) to generate one index to this reference genome. Next, character by character of each read is mapped

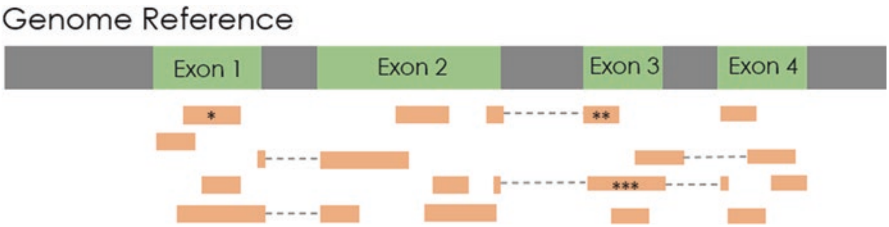


Fig. 3.4 Short reads mapped to a reference genome. Reads are aligned to a reference genome and the accumulation of data brings in evidence expressed exons and splice junctions

until the entire sequence is aligned. If a read cannot find a perfect alignment location, the program backtracks one character, substitutes this character, and the process is repeated until the alignment is completed. The maximum number of character substitutions is a parameter in Bowtie. The rapid improvement of throughput and increase of read length of sequencing technologies required the development of Bowtie2 (Langmead and Salzberg 2012), a gapped supported alignment tool that performs a faster and more sensitive mapping for reads longer than 50 bp.

TopHat (Trapnell et al. 2009) can identify exons splicing sites by mapping RNA-seq reads against a reference genome. First, the Bowtie mapping program is employed to map the short unspliced reads to the reference genome. The reads that are not initially mapped are not filtered out but are just set apart. After the main alignment, each unmapped reads are split into shorter fragments and then aligned individually and independently to identify splice junctions between exons. TopHat2 (Kim et al. 2013) is an updated version of TopHat with an overall accuracy improvement and better alignment procedure.

The Spliced Transcripts Alignment to a Reference (STAR) (Dobin et al. 2013) represents a significant mapping alignment algorithm for RNA-seq data. STAR aligns non-contiguous (exons) sequences straight to a reference genome by two main steps. First, in the seed searching phase, a maximal mappable prefix (MMP) is employed to correctly map the reads against the reference genome even if the read contains a splice junction. Later, the algorithm attaches the seeds previously aligned and constructs alignments of all read sequences. Finally, using a defined local alignment score system, a seed combination is called the best alignment for a read if it has the highest score.

Segemehl (Hoffmann et al. 2009, 2014) maps short reads to reference genomes, detecting mismatches, insertions, and deletions. Moreover, Segemehl can deal with different read lengths and can map primer or polyadenylation contaminated reads correctly. Segemehl matching method is based on enhanced suffix arrays, supporting the SAM format and queries with gzipped reads to save disk and memory space and allowing both bisulfite sequencing and split read mappings.

Minimap2 (Li 2018) is a fast RNA-seq aligner that maps long-reads against a reference database. Minimap deals with long noisy reads at high error rates generated from both ONT and PacBio sequencing. In aligning spliced sequences, it recovers insertions and deletions and predicts correct splice junctions for correct alignment.

There are many other computational methods to map short reads to a reference genome, as shown in Table 3.2.

3.2.2.3 Software for Assembling

Mapping approaches for transcriptome reconstruction can be particularly tricky since correctly assigning reads to a reference genome are usually computational demanding, prone to errors by splice junctions, sequencing inaccuracy, absence, or unfinished reference sequences. Contrarily, assembly (or de novo assembly) approaches do not require any reference genome, the desired feature, especially when genomic sequences are not available or do not attend minimum quality demands.

The assembly tools algorithms usually aim to group reads with similar extremities, i.e., the overlapping of one read's end to another indicates that both probably belong to the same transcript (Fig. 3.5). These similar extremities enable the reconstruction of larger regions of the transcripts. As said before, each of these groups is called a contig. The sequence resulting from the overlapping reads in one contig, called consensus, is a predicted (fragment of) transcript.

Short reads sequencing usually have greater accuracy than long reads; however, short reads often align in multiple regions, causing problems to find correct isoforms. Thus, long reads sequencing can improve the discovery and identification of isoforms, but it is less accurate due to base-calling errors. When possible, the mixture of long reads and Illumina short reads are the best strategy for assembling complete and accurate transcriptomes (Kovaka et al. 2019).

Trinity (Grabherr et al. 2011) software package represents a major de novo assembly method composed of three modular components: Inchworm, Chrysalis, and Butterfly. Initially, the inchworm algorithm decomposes and selects from all reads the most common k-mer ($k = 25$) as the seed promotes contig assembly based on greedy extension ($k-1$)-mer overlaps. Chrysalis clusters and connects Inchworm contigs in components that could be originated from alternative splicing or related

Table 3.2 Mapping software and their websites

Mapping softwares	Website/repository
Bowtie1/Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Minimap2	https://github.com/lh3/minimap2
Segemehl	https://www.bioinf.uni-leipzig.de/Software/segemehl/
STAR	https://github.com/alexdobin/STAR
TopHat2	http://ccb.jhu.edu/software/tophat/index.shtml
NextGenMap	http://cibiv.github.io/NextGenMap/
Kallisto	https://pachterlab.github.io/kallisto/
HPG Aligner	https://github.com/opencb/hpg-aligner

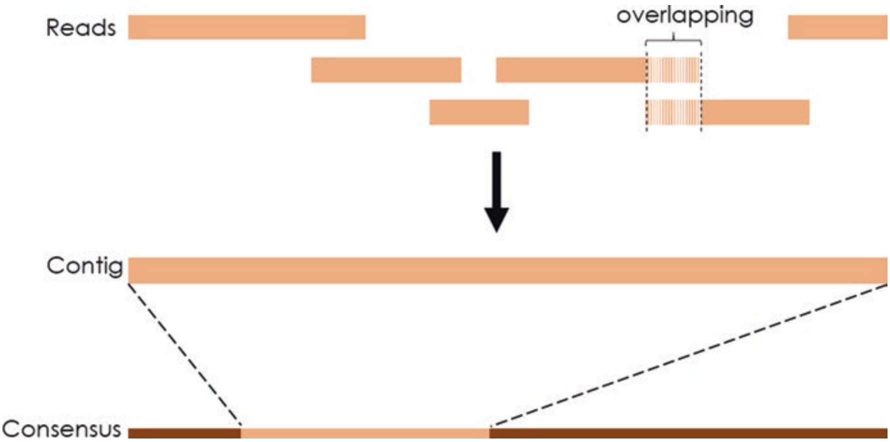


Fig. 3.5 Reads that contain overlapping extremities indicate that they are parts of the same transcript. Multiple reads overlapping each other creates a longer fragment called contig that represents a specific *locus* of consensus sequence

genes. If contigs overlap k−1 bases between themselves and reads span the splicing junction among different contigs, then highly structured de Bruijn graphs are built for each component. Finally, the Butterfly component integrates de Bruijn graphs produced in the Chrysalis stage to their corresponding RNA-seq read, allowing the reconstruction of the transcriptome sequences similar to the original transcripts.

Trans-ABySS (Transcriptome Assembly By Short Sequences) (Robertson et al. 2010) is a de novo assembly tool designed to reconstruct paired-end short reads from transcriptome data. Trans-AbySS derived from ABySS (Simpson et al. 2009), a short-read genomic data assembler. Trans-AbySS employees de Bruijn graph approach promoting data assembly with standard k-mers (k = 32) promoting a good balance between assembling frequent and rare transcripts. Trans-ABySS single-processor version is useful for assembling genomes of up to 100 Mbases. In contrast, the parallel version (implemented using MPI) can be assembled larger genomes, benefiting from multi-threaded processing.

MaSuRca (Zimin et al. 2013) process hybrid assembly, using “super-reads” from short-reads to de novo assemble reads and construct synthetic long reads with a low error rate and combining with long reads from Nanopore/Pacbio. Its assembly permits work with long reads and short reads at the same time, overcoming high error rates from long-reads sequencing (Table 3.3).

3.2.2.4 Software for Analysis

In transcriptome projects, quantitative analysis, differential expression, and transcript annotation are extensively used. Many suitable tools for these analyses are available in R language, which provides a wide variety of statistical and graphical

Table 3.3 Assembly software and their websites

Assembly	Website/repository
BWA	https://github.com/lh3/bwa
Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/
MaSuRca	https://github.com/alekseyzimin/masurca
SPAdes	http://cab.spbu.ru/software/spades/
StringTie2	https://github.com/skovaka/stringtie2
Trans-ABySS	https://github.com/bcgsc/transabyss
Trinity	https://github.com/trinityrnaseq/trinityrnaseq/wiki
SOAPdenovo	https://github.com/aquaskyline/SOAPdenovo-Trans
Oases	https://github.com/dzerbino/oases

resources. R is highly extensible, allowing us to output well-designed publication-quality plots, including effective data handling and storage facility and a collection of intermediate tools for data analysis. Bioconductor (https://www.bioconductor.org) is a (mostly) R packages repository that provides open-source tools to analyze biological high-throughput data. Similarly, there are many Python-based resources as Biopython (https://biopython.org), a set of freely available tools for biological computation written in Python.

Quantitative Analysis

The transcript coverage is the number of reads "covering" (or the number of mapped reads in) a transcript. The greater the number, the more abundant is the expressed gene in an RNA-seq sample (Fig. 3.6). The RNASeqMap library (Leśniewska and Okoniewski 2011), for instance, provides classes and functions to analyze the RNA-sequencing data using the coverage profiles in multiple samples at a time.

Differential Expression

The differential expression refers to the study of the variability of genetic expression between samples. One important objective of RNA-seq projects is to identify the differentially expressed genes in two or more conditions (Rapaport et al. 2013). These genes are selected based on parameters, usually based on p-values generated by statistical modeling. The expression level is measured by the number of reads mapping to the transcript, such as transcripts per million (TPM), which is expected to correlate directly with its abundance level. This measure is different from gene probe-based methods, e.g., microarrays. In RNA-seq, the expression of a transcript is limited by the sequencing depth. It depends on the expression levels of other transcripts, in contrast to array-based methods, in which probe intensities are independent of each other. That one and other technical differences have motivated many statistical algorithms, with different approaches for normalization and

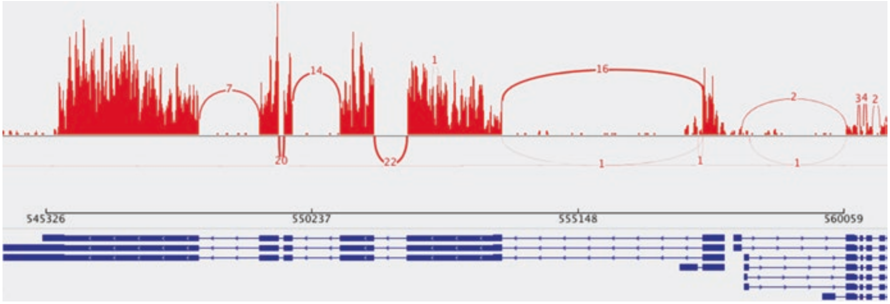


Fig. 3.6 Read coverage of transcripts relative to a reference genome. Each red bar plotted indicates a *locus* alignment coverage. The arcs represent splicing junctions between exons. Finally, the arc numbers are the observed numbers of reads across the junction. (Source: https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html)

differential expression detection. For example, Poisson or negative binomial distributions to model the gene count data and various normalization procedures are common approaches.

Cufflinks (Trapnell et al. 2010) may be used to measure global de novo transcript isoform expression. It assembles transcripts, estimates their abundances, and determines differential expression (Trapnell et al. 2013) in RNA-seq samples. Moreover, Cufflinks accepts reads aligned by other mappers and assembles the alignments to a parsimonious set of transcripts. It then estimates the relative abundances of these transcripts based on how many reads support each one, considering biases in library preparation protocols.

Some articles discuss and compare statistical methods to compute differential expression. In a review, Kvam et al. (2012) compared four statistical methods – edgeR, DESeq, baySeq, and a method with a two-stage Poisson model (TSPM). Rapaport et al. (2013) describe an extensive evaluation of common methods – Cuffdiff (Trapnell et al. 2013), edgeR (Robinson et al. 2010), DESeq (Anders and Huber 2010), PoissonSeq (Li et al. 2012), baySeq (Hardcastle and Kelly 2010), and limma (Smyth 2004) adapted for RNA-seq use, using the Sequencing Quality Control (SEQC) benchmark dataset and ENCODE data.

Splice Junctions

Splice junctions are nucleotide sequences at the exon–intron boundary in the pre-messenger RNA of eukaryotes removed during the RNA splicing. This process can generate many processed transcripts from a single gene. Computationally, the problem is to recognize, given a sequence of DNA, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). This problem consists of two subtasks: recognizing exon/intron boundaries (called EI sites) and recognizing intron/exon boundaries (IE sites). IE borders are called “acceptor sites,” while EI borders are called “donor sites.” The recognition and quantification of splice variants are among the advances of RNA-seq over microarray to measure differential gene expression.

The splice junctions help to delineate and quantify the transcript model, as observed in Fig. 3.6.

Tophat (Trapnell et al. 2009) identifies splice junctions, producing the junctions. bed file, where the field score is used to indicate coverage depth. The identified splice junctions can be displayed in browsers (e.g., UCSC genome browser (Kuhn et al. 2013)) using.bed files encoding splice junctions. Junction files should be in the standard.bed format. Pasta (Patterned Alignments for Splicing and Transcriptome Analysis) (Tang and Riva 2013) is a splice junction detection algorithm designed for RNA-seq data, based on a highly accurate alignment strategy on a combination of heuristic and statistical methods to identify exon–intron junctions with high accuracy.

Annotation

The annotation step aims to assign a biological function for each transcript, identifying genes and finding more information, e.g., biological categories and ontologies. The annotation process is characteristic of novel transcriptomes since reference genomes and transcriptomes are typically associated with curated gene annotation.

The annotation methods can be organized into two classes:

- Pairwise comparison of every transcript against a file with known transcripts and their corresponding annotation. This can be done by comparing the nucleotides or the translated nucleotides.
- *Ab initio* gene prediction, where the presence of structural features and motifs of known genes are used to infer function.

The pairwise sequence comparison (or pairwise alignment), where a query sequence (transcript of the organism of interest) is compared with annotated sequences datasets, relies on an algorithm that computes an alignment among two transcripts. The hypothesis is based on Darwin's evolution theory, which claims that living organisms evolved from ancestor organisms. Therefore, if two transcripts have similar sequences, they may be homologs and probably share the same biological functions. This means that biological function may be inferred from similar sequences. Important pairwise algorithms, which produce alignments between pairs of sequences, are Smith-Waterman (Smith and Waterman 1981) and BLAST (Altschul et al. 1990).

Similar to the assembly step, the main difficulty in the annotation is due to the transcript length. The resulting genes may be fragmented, causing loss of information. Since alignment programs are error-tolerant, it is reasonable to expect that the annotation for transcripts (predicted from reads generated by high-throughput sequencers) is correct if functions of genes of other organisms have been found correctly.

In contrast, finding genes ab initio is not so error robust since sequencing errors can lead to incorrect gene prediction. In particular, sequencing errors introducing a stop codon can result in an incorrectly predicted gene.

3.3 Single-Cell Transcriptome Sequencing (scRNA-seq)

Although cells in an organism share almost identical genotypes, gene expression is heterogeneous and reflects the activity of a subset of genes. ScRNA-seq technologies are capable of generating data sets that describe the transcriptome of single cells. Single-cell transcriptome sequencing (scRNA-seq) expands the biological panorama granted by RNA-seq. It allows to estimate the expression levels of the whole transcriptome or targeted gene expression from a single cell and addresses new biological questions such as the heterogeneity of cell responses and their gene regulatory networks. It emerged with an mRNA-seq assay where a single mouse blastomere was sequenced, detecting the expression of 75% more genes than microarray techniques (Tang et al. 2009). This pioneer scRNA-seq method profiled RNA transcriptomes from single cells using oligo-dT primers followed by ligation adapter PCR (Tang et al. 2009). This method's limitation is the reverse transcriptase's inefficiency on the first-strand cDNA synthesis, causing a 3′ bias.

Eventually, new protocols and lower sequencing costs made scRNA-seq more accessible as technologies advance, resulting in continuously growing datasets, ranging from ~10^2 to ~10^6 cells. Some of the most distinguished methods for scRNA-seq are Smart-seq (Ramsköld et al. 2012), Smart-seq2 (Picelli et al. 2014), Drop-seq (Macosko et al. 2015), inDrop (Klein et al. 2015), CEL-seq2 (Hashimshony et al. 2016), 10× Chromium (Zheng et al. 2017), and Smart-seq3 (Hagemann-Jensen et al. 2020).

In general, scRNA-seq methods tag transcripts to make it possible to identify their cell of origin and generate libraries for sequencing. scRNA-seq sequencing data can both come from next-generation sequencing (NGS) and single-molecule sequencing (SMS) (Gao 2018). Smart-seq, Smart-seq2, Smart-seq3, and CEL-seq2 can be considered low-throughput plate-based methods, where the cells are sorted into wells of a multi-well plate. Alternatively, bead-based high-throughput methods distribute the cell suspension into tiny droplets containing reagents and barcoded beads (Drop-seq, 10× Chromium, and inDrops) or into well microplates (Seq-Well and sci-RNA-seq) to produce single droplets or well microplates with one cell and one bead marking the cDNA generated from that cell (Ding et al. 2019).

The Smart-Seq (Ramsköld et al. 2012) addressed this problem using a Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT) to synthesize cDNA with long messenger RNA templates. Unique molecular identifiers (UMI) were incorporated into each RNA molecule as unique barcodes before the whole transcriptome amplification (WTA) amplification (Islam et al. 2014). Smart-seq2 (Picelli et al. 2014) is an approach that combines sensitivity (it captures a considerable fraction of RNAs present in cells) with full-length coverage of transcripts and can detect genes per cell and across cells enabling quantifying isoform-level expression from single cells, but without the incorporation of unique molecular identifiers (UMIs). Smart-seq3 (Hagemann-Jensen et al. 2020) improves the sensitivity of Smart-seq2, adding optimized reverse transcriptase and buffer conditions together with a partial Tn5 motif and a tag sequence in the template-switching oligonucleotide to directly assign individual RNA molecules to isoforms and establish their allelic origin in single cells.

Drop-Seq dissociates a tissue into individual cells and encapsulates them into droplets with microparticles that deliver barcoded primers. After associating barcodes to each cell's RNAs, they are reverse-transcribed into cDNAs to generate beads called "Single-cell Transcriptomes Attached to Microparticles" (STAMPs). Then, the STAMPs are amplified in pools for high-throughput mRNA-seq (Macosko et al. 2015) (Fig. 3.7). The 10× Chromium system works, generating a large number of "Gel Bead-in-emulsions partitions" (GEMs) to index each cell's transcriptome separately. The barcoded gel beads (read, 10xbarcode, UMI, oligo-dT) are mixed with cells, enzymes, and partitioning oil to create single-cell GEMs. Then, the single-cell GEMs undergo reverse transcriptase (RT) to generate a 10× barcoded cDNA library where cDNA from individual cells share a common 10× barcode that can be used for single-cell whole transcriptome sequencing or target sequencing workflows (10× Genomics Inc. 2020). In the inDrops method, the cells are also encapsulated into droplets with lysis buffer, hydrogel microspheres carrying barcoded primers, and an RT mix. After the release of primers, cDNA in each droplet is barcoded during reverse transcription. After the droplets are broken, all cellular material can be amplified for sequencing (Klein et al. 2015).

The Smart-seq methods can detect many genes in a cell, including low abundance transcripts and alternatively spliced transcripts. CEL-seq2 (Hashimshony et al. 2016), Drop-seq, 10× Chromium, and inDrops can quantify mRNA levels with less amplification noise using UMIs, enabling less and profiling isoform-level RNA counting. As a limitation, inDrops droplets may contain two cells or two different types of barcodes. Table 3.4 shows a comparison of some important aspects of these scRNA-seq methods.

3.3.1 scRNA-seq Computational Analysis

Despite the different methods available, the scRNA-seq data is essentially the result of high-throughput sequencing cDNA reverse transcribed from mRNA isolated from a pool of cells. The primordial difference is that the sequenced data is somehow tagged to assign its origin to individual cells. Some standard steps remain the same as RNA-seq, such as the reads quality filtering and reads mapping to a reference genome. Reads quality filtering can be applied to filter the read quality using a quality metric for sequencing like the percentage of base calls (Q score). The reads are then mapped to a reference genome and quantified to generate an expression profile matrix. Some scRNA-seq specialized tools can both align and quantify the reads. Additionally, a second filtering step can be performed after quantifying reads to discard cells expressing a low number of genes or a high number of mitochondrial genes (Park and Lee 2020). The next step of the pipeline is data normalization using a metric for expression normalization as TPM (Transcripts Per Kilobase Million) or RPKM (Reads Per Kilobase per Million) (Gao 2018). At this point, the scRNA-seq computational analysis reaches its two fundamental problems: cluster analysis and sample/feature reduction.

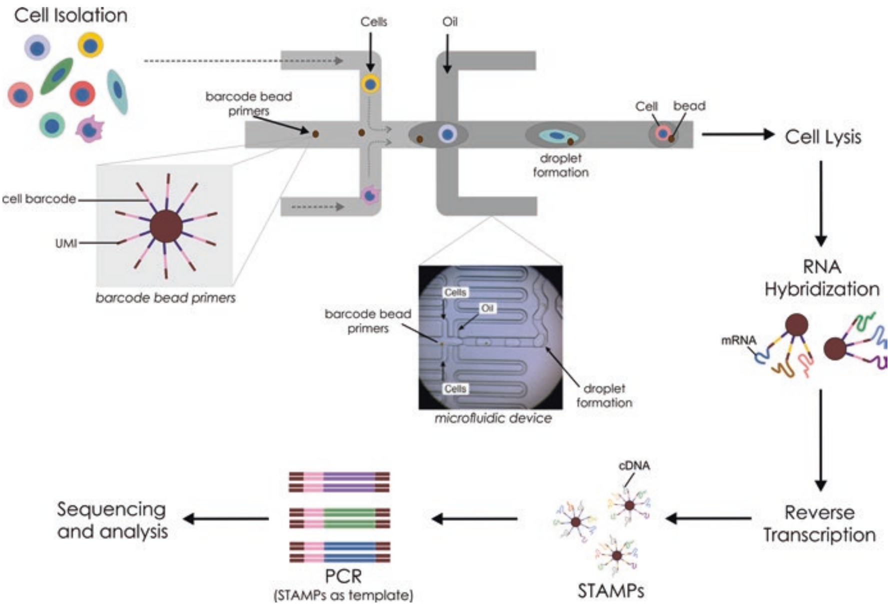


Fig. 3.7 Individual cell's transcriptome can be analyzed using scRNA-seq. Tissue disrupted single cells are mixed with barcode bead primers and reagents in oil droplets in a microfluidic device. The formed droplet contains a single cell and a barcode. After lysis and primer hybridization, RNA is reverse transcribed and sequenced as in a conventional RNA-seq experiment. The UMI and barcode sequence will be incorporated in the final sequenced reads and will guide the scRNA-seq processing

Normalization allows consistent comparison of gene expression measurements in individual cells, including technical variation due to the numbers of sequenced readings or transcripts identified per cell. A normalized gene expression matrix is a matrix with n samples (cells) by m features (genes, transcripts, or exons), depending on the read's size. For example, for transcripts as features, PacBio full-length transcriptome could be the right choice, or for Illumina short-length reads, the features could be genes. As the number of annotated genes of the target organism, the matrix could be large and sparse, which justifies the sample and feature reduction. The feature selection can be understood as removing genes unhelpful to distinguish biological variation across samples.

Clustering cells allow us to identify cells with correlated phenotype by grouping them based on their gene expression profiles' similarity. This is achieved using dimension reduction algorithms to embed the expression matrix into a low-dimensional space that summarizes the data structure in as few dimensions as possible (Gao 2018; Luecken and Theis 2019). These low-dimensional spaces can come from dimension reduction methods as Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), Multidimensional Scaling (MDS), and t-distributed Stochastic Neighbor Embedding (t-SNE).

Table 3.4 Comparison of some aspects of low- and high-throughput scRNA-seq methods

	Low-throughput				High-throughput		
Methods	Smart-Seq (Ramsköld et al. 2012)	Smart-seq2 (Picelli et al. 2014)	Smart-seq3 (Hagemann-Jensen et al. 2020)	CEL-seq2 (Hashimshony et al. 2016)	Drop-seq (Macosko et al. 2015)	10× Chromium (Zheng et al. 2017)	inDrop (Klein et al. 2015)
Single-cell isolation	Micromanipulation	[a]MCP/FACS	FACS	FACS	Droplets	Droplets	Droplets
Coverage	Full length	Full length	Full length	–	–	–	–
UMI	No	No	No	6 bp	8 bp	28 bp (Cell barcode & UMI)	6 bp
First-strand synthesis	oligo-dT	oligo-dT	oligo-dT	oligo-dT	oligo-dT	oligo-dT	oligo-dT
Second-strand synthesis	Template switching	Template switching	Template switching	RNAseH/DNA Pol	Template switching	Template switching	Template switching
cDNA Amplification	PCR	PCR	PCR	IVT	RT-PCR	RT-PCR	RT-PCR

[a] Microcapillary Pipette (MCP)

3.3.2 scRNA-seq Analysis Tools

Seurat (Hao et al. 2020) is an R package that integrates quality control, analysis, and exploration of single-cell RNA-seq data. It is based on a *Seurat object*, which serves as a container for both data (like the count matrix) and analysis (like PCA, or clustering results). Also, Seurat can make simultaneous measurements of multiple data types from the same cell, known as multimodal analysis, and analyze spatially resolved RNA-seq data.

Cell Ranger is a set of tools to process Chromium single-cell RNA-seq data. The package contains *cellranger mkfastq* which demultiplexes raw base call (BCL) Illumina files into fastq files. These files are then taken as input by *cellranger count* to perform alignment, filtering, barcode, and UMI counting. In the next step, *cellranger aggr* aggregates and normalizes the outputs from multiple runs of *cellranger count* recomputing the feature-barcode matrices and analyzing the combined data. The *cellranger reanalyze* reruns the dimensionality reduction, clustering, and gene expression algorithms from the feature-barcode matrices produced by *cellranger count* or *cellranger aggr*. Cell Ranger also uses the aligner STAR (Dobin et al. 2013) and the output is delivered in formats like bam, mex, csv, hdf5, and html.

Meta Cell (Baran et al. 2019) is a tool for deriving metacells and analyzing scRNA-seq data. Metacells are a theoretical group of scRNA-seq cell profiles statistically equivalent to samples derived from the same RNA pool, which is obtained by computing partitions of scRNA-seq datasets into disjoint and homogenous groups of cells.

SEQC (Azizi et al. 2018) is a Python package for scRNA-seq analysis in a cloud and subsequent analyzes on a local machine. It has Spliced Transcripts Alignment to a Reference – STAR (Dobin et al. 2013), Samtools (Li et al. 2009), and HDF5 data model as dependencies and has been tested for 10× Genomics v2 and inDrop v2 data.

zUMIs (Parekh et al. 2018) is a pipeline to process RNA-seq data with or without UMIs. zUMIs take cDNA fastq files and other reads containing UMI and Cell Barcode information as input. It was written using R, Perl shell, and Python programming languages and has as dependencies STAR (Dobin et al. 2013).

robustSingleCell is an R package that provides clustering and comparison of population compositions across tissues and experimental models through a similarity analysis characterizing transcriptomic similarities in meta-clusters by identifying their defining overexpressed genes (Magen et al. 2019) (Table 3.5).

Table 3.5 Computational tools for scRNA-seq analysis

Tools	Availability
Seurat (Hao et al. 2020)	https://github.com/satijalab/seurat
SEQC (Azizi et al. 2018)	https://github.com/ambrosejcarr/seqc
zUMIs (Parekh et al. 2018)	https://github.com/sdparekh/zUMIs
CellRanger (10× Genomics)	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger
Meta Cell (Baran et al. 2019)	https://tanaylab.github.io/metacell
robustSingleCell (Magen et al. 2019)	https://github.com/asmagen/robustSingleCell

3.4 Case Study 1

RNA-seq as an Efficient Tool to Analyze and Identify Gene Expression Patterns Related to Murine Bone Marrow-Derived Macrophage's Susceptibility and Resistance to Candida albicans *Infection*

The improvements in organ transplantation techniques and the rise of immune-compromised diseases, like AIDS, are directly linked to the exponential growth of opportunist infections in these patients. Therefore, the study of the etiological agents of these diseases, particularly fungal pathogens, together with the immune response they elicit, became an important issue (Marr et al. 2002; Richardson and Lass-Flörl 2008; Miceli et al. 2011). *Candida albicans* appears to be the leading cause of invasive infections among fungi, showing high morbidity and mortality rates (Chi et al. 2011; Shigemura et al. 2014).

Many studies have been done to understand the aspects of immune responses to *C. albicans* (Tierney et al. 2012; Miramón et al. 2013; Hünniger et al. 2014; Martínez-Álvarez et al. 2014). In this case study, the transcriptomic response of murine bone marrow-derived macrophages (BMDMs) from BALB/c (resistant) and DBA/2J (susceptible) mice strains to *C. albicans* infection was analyzed by RNA-seq to compare both transcriptomic patterns. Therefore, this case study's main objective was to identify BMDMs gene expression patterns between resistant and susceptible mice after *C. albicans* infection by the analysis of the resulting transcriptome profiles.

Bone marrow was extracted from the mice, and the hematopoietic stem cells were then differentiated into macrophages. An amount of 2 × 10^6 BMDMs were co-cultured with 4 × 10^6 *C. albicans* yeasts for 90 min, and the RNA was extracted using RNeasy (Qiagen). RNA quality and concentration were verified employing a Bioanalyzer (Agilent) and NanoDrop (Thermo Scientific), respectively. Three μg of total RNA was used for the library preparation, including a step of rRNA depletion using Ribozero (Epicentre) before library construction and sequencing in an Illumina Hiseq platform.

The sequencing results were provided in fastq format. FastQC was used to assess quality. Adaptors clipping and quality trimming were performed using Cutadapt

(Martin 2011). Two mapping software, NextGenMap (NGM) (Sedlazeck et al. 2013) and Tophat2 (Kim et al. 2013), were employed. Since both generate a similar number of mapped reads, we chose NextGenMap due to its faster analysis. Low-quality mappings were removed using Samtools (Li et al. 2009), which was also used to sort, index, and convert the mapping results from sam to.bam files. Bedtools (Quinlan and Hall 2010) were then used to count reads for both genes and exons, and generate a table of these counts, to be analyzed for differential expression. As said before, differential expression can be analyzed using different methodologies (Wagner et al. 2012; Soneson and Delorenzi 2013), and EdgeR (Robinson et al. 2010) and DESeq (Anders and Huber 2010) were chosen. Both outputted very similar results. Alternative splicing can be checked by differential exons usage (Anders et al. 2012). Therefore, the resulting list of genes or transcripts differentially expressed (adjusted *p*-value <0.05 and fold change ≥ ±2.5) was checked for gene ontology (GO terms) using ClusterProfiler (Yu et al. 2012) Bioconductor package.

Several problems may occur in RNA-seq projects, and here we point out some of these:

- Infection conditions: the optimization of the protocols of co-culture conditions, as well as RNA extraction, may be hard to adjust. Setting a multiplicity of infection (MOI – proportion of host/pathogen cells in the co-culture) that suffices to induce a transcriptomic response in the host cells is the first step. However, a very high MOI may result in host cells' death and apoptosis, which may result in altered gene expression or low amounts of RNA extracted from these cells.
- Infection time: the definition of correct time intervals of interaction between pathogen and host cells is essential since different genes have different kinetics of transcription during co-culture. This may vary drastically for different host-pathogens and also depends on the major question of interest.
- Biological replicates: in transcriptomic studies, robust statistical analysis is fundamental. In this sense, the experimental design has to incorporate proper biological replicates to allow valid statistical inferences (Robles et al. 2012).
- Library preparation and sequencing parameters: the choice of the preparation methodologies, e.g., poly-A enrichment protocols versus rRNA depletion protocols, or paired-end versus single-end sequencing, may strongly impact the results. Improper handling of samples in this step may also result in sample degradation or inefficient rRNA depletion, which may compromise the whole experiment if not properly adjusted. A well-defined experimental design for the sequencing step must also be taken into consideration. A final low coverage of the transcriptome can result in an inadequate analysis of differential gene expression.

A significant disparity was observed in the differentially expressed genes upon *C. albicans* infection between BMDMs from both mice strains. BMDMs from the susceptible DBA/2J strain modulated a higher number of genes (4021) upon infection with *C. albicans* than BMDMs from the resistant BALB/c strain (99), and both sets have few genes in common (60) (Fig. 3.8).

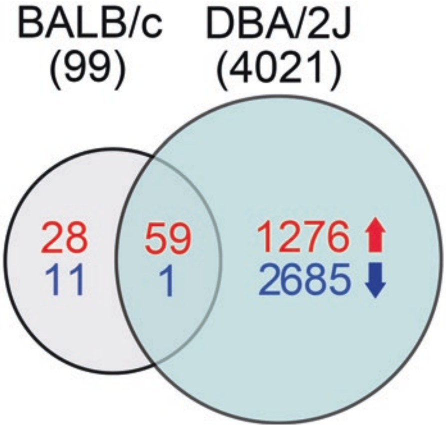


Fig. 3.8 Venn diagram of positively (red) and negatively (blue) regulated genes in BMDMs from BALB/c and DBA/2J mice strains infected with *C. albicans*. Differentially expressed genes were considered when adjusted *p*-value <0.05 and fold change ≥±2.5

Analysis focusing on GO categories of biological processes revealed enrichment (p <0.01) of upregulated genes in terms related to inflammatory response, cellular response to biotic stimulus, and cytokine production in both resistant and susceptible strains (Fig. 3.9). However, they markedly differed in the modulation of some terms. For example, macrophages from the resistant strain upregulated genes related to apoptosis and neutrophil chemotaxis. In contrast, macrophages from the susceptible strain upregulated genes involved in innate immune response and leukocyte migration.

3.5 Case Study 2

Single-Cell Sequencing of SARS-CoV-2 Infected Individuals with Distinct Levels of Severity

COVID-19 outbreak has caused critical consequences for all countries, including many deaths and hospitalization, beyond the economic issues. Beyond the vaccination, it is important to research specific drugs to treat the affected individuals. Monoclonal antibodies have demonstrated their effectiveness in medicine (Maranhão et al. 2020). Therefore, developing new potential antibodies as an alternative against viral proteins remains highly valuable.

This example of scRNA-seq analysis is based on the work “Single cell RNA and immune repertoire profiling of COVID-19 patients reveals novel neutralizing antibody” from Fang Li et al. (2020). They have conducted a study using single-cell transcriptome sequencing (scRNA-seq), single-cell BCR sequencing (scBCR-seq), and deep BCR repertoire to reveal neutralizing antibody sequences in patients who have recently cleared the virus. They collected blood samples (peripheral blood mononuclear cells – PBMCs) from 16 COVID-19 patients and eight healthy controls to reveal immune cells’ changes caused by SARS-CoV-2 infection. Fang Li et al. (2020) scRNA-seq was performed using 10× Genomics. The original data is available in the Zenodo under the accession URL: https://zenodo.org/record/3744141.

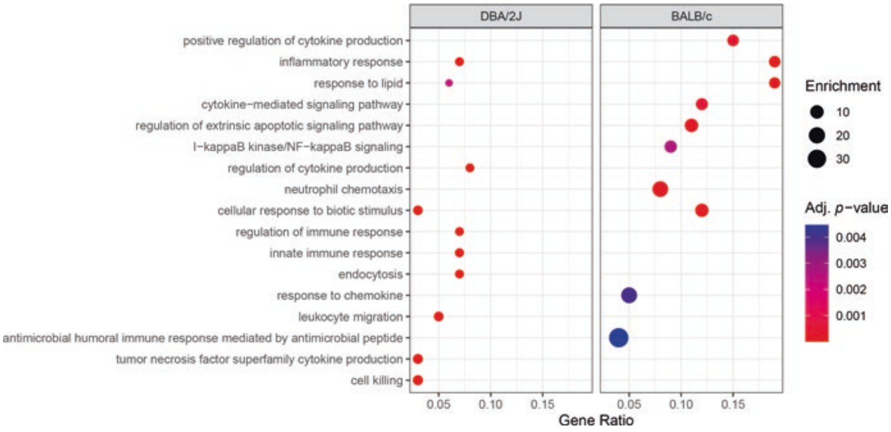


Fig. 3.9 Gene ontology enrichment of upregulated genes in BMDMs from DBA/2J and BALB/c mice strains upon *C. albicans* infection. Enriched GO terms (adjusted *p*-value <0.01) from biological processes category associated with upregulated genes in BMDMs derived from the susceptible DBA/2J (left) and the resistant BALB/c (right) mice strains. Dot size is representative of enrichment (gene modulated ratio/gene background ratio) for each GO term. Only major terms related to immune response were plotted

This case study uses a Fang Li et al. (2020) sample subset with data from two patients to demonstrate how to identify distinct types of cells based on clustering their transcripts and how to obtain the differentially expressed genes. The input files are barcodes.tsv, datasets.rds, genes.tsv, and matrix.mtx. For this case study, we filtered the complete data to work only with patient 3 (P3) and patient 10 (P10) samples, both from 59 years old females with distinct levels of COVID-19 severity. P3 had severe symptoms, and P10 had moderate symptoms.

This example uses the R package Seurat 4.0 (Hao et al. 2020) to perform the analysis directly from the matrix. The following R codes are commented, and their results presented. The first step is to install and load the required R packages. Seurat 4.0 requires R version 4.x.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.12")

BiocManager::install('ggplot2')
BiocManager::install('ggrepel')
BiocManager::install('limma')
BiocManager::install('calibrate')
BiocManager::install('dplyr')
BiocManager::install('Matrix')
BiocManager:: install('Seurat')

library(ggplot2)
library(ggrepel)
library(limma)
library(calibrate)
library(dplyr)
library(Matrix)
library(Seurat)
```

The next step is to download, extract, and read the COVID-19 data. This will result in a matrix with 33,538 lines and 96,404 columns. The columns represent each tagged transcript, and the lines represent the genes where those transcripts were mapped.

```
system("wget https://zenodo.org/record/3744141/files/COVID-19.tar.gz")
system("tar -xzvf COVID-19.tar.gz")
covid_19_data <- Read10X(data.dir = "COVID-19")
dim(covid_19_data) # dimensions for full data
```

Once loaded the full data, now it is possible to filter them to work only with P3 and P10 samples by using regular expression to identify only data from patients P3 and P10. The new dimensions of P3 and P10 data will be 33,538 lines (genes) by 16,056 columns (tagged transcripts).

```
p3_and_p10_data <- covid_19_data[, grep(pattern = "P3|P10", colnames(covid_19_data))]
dim(p3_and_p10_data) # dimensions for selected data
```

The function CreateSeuratObject() initializes the Seurat object with the non-normalized data constrained by the following parameters: minimal of two cells with at least 20 expressed genes and at least 2,000 features. The dimension of the object in this case will be 17,169 genes and 2,123 tagged transcripts that met the criteria.

```
covid_p3_p10 <- CreateSeuratObject(
    counts = p3_and_p10_data,
    project = "COVID-19",
    min.cells = 2,
    min.genes = 20,
    min.features = 2000
)
dim(covid_p3_p10) # dimensions for loaded data
```

Before starting the data processing, we will create two new columns to add meta-information for the patients (P3 or P10) and for the mitochondrial percent in transcripts. The [[]] operator can add columns to an object. In this case, we create a column to identify patients P3 and P10. We also stash quality control (QC) stats for their mitochondrial samples, which are identified starting by "MT-".

```
covid_p3_p10[["patient"]] <- sapply(strsplit(colnames(covid_p3_p10),"-"), `[`, 1)
covid_p3_p10[["perc_mitochondrial"]] <- PercentageFeatureSet(covid_p3_p10, pattern = "^MT-")
```

Next, it is possible to build a violin plot to visualize the QC metrics for number of features, read count and mitochondrial percentage, grouped by patient (Fig. 3.10).

```
plot_perc_mitochondrial <- VlnPlot(
    covid_p3_p10,
    features = c("nFeature_RNA", "nCount_RNA", "perc_mitochondrial"),
    ncol = 3,
    Group.by = "patient",
    log = TRUE
)
plot_perc_mitochondrial
```

The next step is to remove unwanted cells from the dataset. In this case we can apply a new filter to keep only samples with the number of features at least equal to 2000 and less than 5% of mitochondrial samples.

```
covid_p3_p10 <- subset(covid_p3_p10, subset = nFeature_RNA >= 2000 & perc_mitochondrial < 5)
```

To normalize the data, we can use function LogNormalize(), which normalizes the feature expression measurements for each cell by the total expression. It multiplies this by a scale factor (10,000 by default), and log-transforms the result.

```
covid_p3_p10 <- NormalizeData(covid_p3_p10, normalization.method = "LogNormalize", scale.factor = 10000)
```

Once normalized, the next step is to identify highly variable features (feature selection) using the method *vst* which, according to the manual of Seurat, fits a line to the relationship of log (variance) and log (mean) using local polynomial regression (loess). Then, it standardizes the feature values using the observed mean and expected variance (given by the fitted line). Then, it is computed the feature variance on the standardized values after clipping to a maximum (default is "auto" which sets this value to the square root of the number of cells).

```
covid_p3_p10 <- FindVariableFeatures(covid_p3_p10, selection.method = "vst", nfeatures = 2000)
```

At this point, it is possible to find, for instance, the 20 most highly variable genes identified (Fig. 3.11) that would be: 'IGHA1', 'JCHAIN', 'IGHG1', 'IGKC', 'IGLC2', 'IGHG2', 'DERL3', 'IGLL5', 'IGHV3-23', 'ITM2C', 'IGKV3-20', 'MZB1', 'LILRA4', 'IGHV3-7', 'FKBP11', 'GNLY', 'IGKV4-1', 'TNFRSF17', 'STMN1', and 'HIST1H4C'. Interestingly, most of these genes are involved with the immune system, more precise to B lymphocytes, a known player in the inflammatory aspect of COVID-19. IGHA, the heavy constant chain of the immunoglobulin alpha, codes for an antibody isotype well characterized to participate in the mucosal immunity, the natural site of SARS-CoV-2 infection.

```
top20 <- head(VariableFeatures(covid_p3_p10), 20)
plot_top20 <- VariableFeaturePlot(covid_p3_p10)
plot_top20 <- LabelPoints(plot = plot_top20, points = top20, size = 2, hjust = .75, vjust = .75)
plot_top20
```

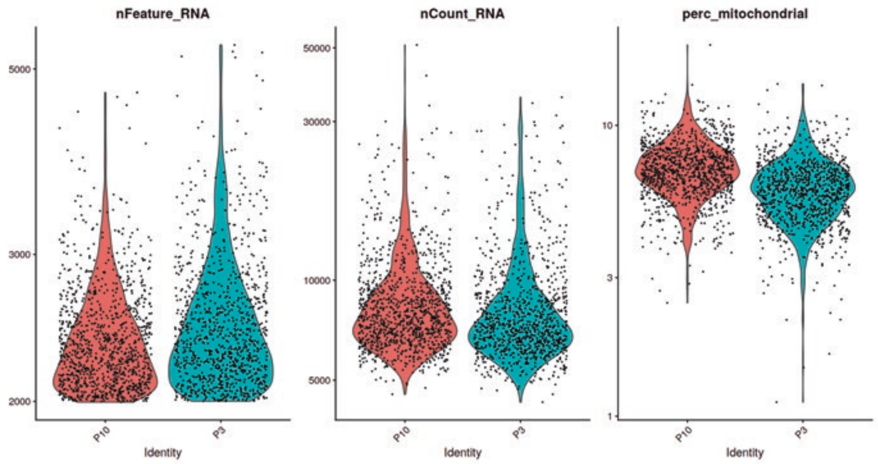


Fig. 3.10 Quality control (QC) metrics for the number of features, read count, and mitochondrial percentage, grouped by patient. Left: Number of featured genes for patients 3 (red) and 10 (blue) after filtering >2000 features. Middle: reads count for P3 and P10. Right: amount of reads from mitochondrial origin shown as percentage

Before performing the dimensional reduction, it is necessary to perform a linear transformation scaling the data. It is a standard pre-processing step prior to applying techniques like PCA.

```
all_genes_covid_p3_p10 <- rownames(covid_p3_p10)
covid_p3_p10 <- ScaleData(covid_p3_p10, features = all_genes_covid_p3_p10)
covid_p3_p10 <- RunPCA(covid_p3_p10, features = VariableFeatures(object = covid_p3_p10))
```

Now, it is possible to determine the dimensionality of the dataset. The function JackStraw() determines the statistical significance of PCA scores by randomly permuting a subset of data, and calculates projected PCA scores for these “random” genes. The ScoreJackStraw() function computes the scores significance by PCs showing a *p*-value distribution that is strongly skewed to the left compared to the *null* distribution.

```
covid_p3_p10 <- JackStraw(covid_p3_p10, num.replicate = 100)
covid_p3_p10 <- ScoreJackStraw(covid_p3_p10, dims = 1:5)
```

We can now cluster the cells. The function FindNeighbors() computes the k.param nearest neighbors for a given dataset using the k-nearest neighbors algorithm. Then, the function FindClusters() identifies clusters of cells from the SNN graph (result of the k-nearest neighbors algorithm). As higher is the resolution parameter, as larger will be the communities.

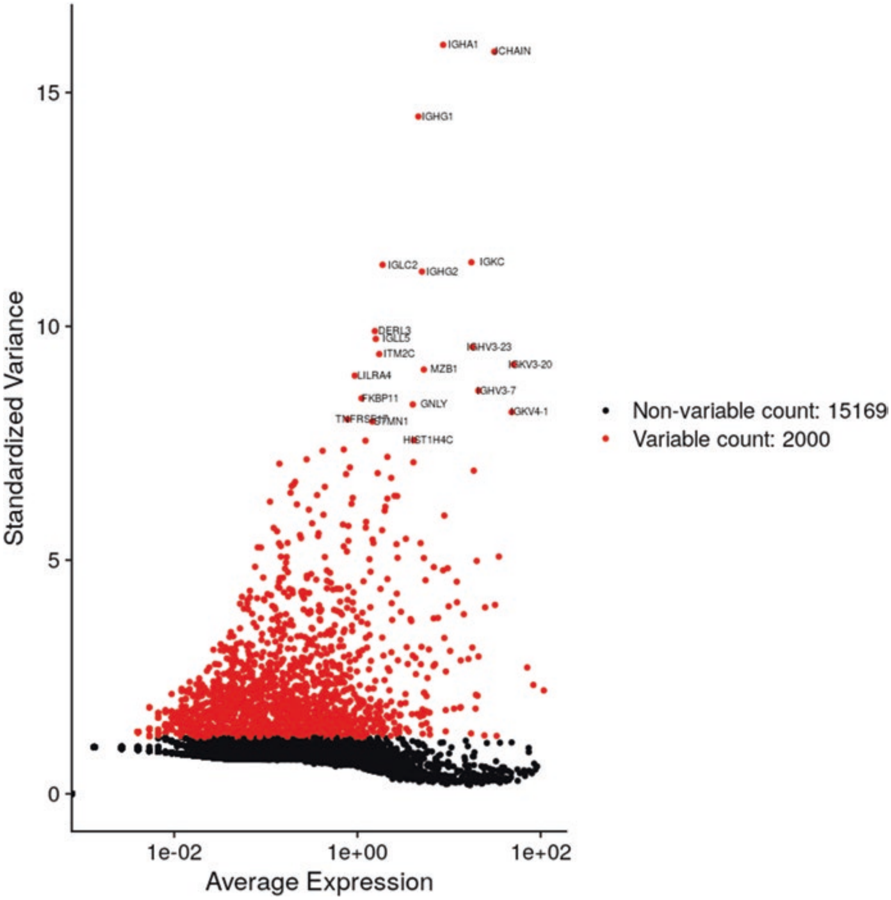


Fig. 3.11 Twenty most highly variable genes identified versus their average expression. In red are shown the 2000 most variable genes among cells, and 20 of them are labeled for exploration purposes

```
covid_p3_p10 <- FindNeighbors(covid_p3_p10, dims = 1:5)
covid_p3_p10 <- FindClusters(covid_p3_p10, resolution = 1)
```

Uniform Manifold Approximation and Projection (UMAP) is a dimensional reduction technique that can be used for visualization similarly to t-SNE, but also for general non-linear dimension reduction. It is founded on three assumptions about the data: (i) the data is uniformly distributed on a Riemannian manifold; (ii) the Riemannian metric is locally constant (or can be approximated as such); and (iii) the manifold is locally connected.

```
covid_p3_p10 <- RunUMAP(covid_p3_p10, dims = 1:5)
# It could be alternatively done using tSNE
# covid_p3_p10 <- RunTSNE(object = covid_p3_p10, dims.use = 1:5)
```

Finally, it is possible to plot the clusters of distinct types of cell in the samples. Using these parameters, we can find 10 clusters as can be seen in Fig. 3.12.

```
plot_clusters <- DimPlot(covid_p3_p10, label = TRUE)
plot_patient <- DimPlot(covid_p3_p10, label = TRUE, group.by = "patient")
plot_clusters + plot_patient
```

As it is possible to see in Fig. 3.12, the cluster number 4 has expressed genes both from patients 3 and 10. In this case, we first split data of patient 3 and 10 and then execute the function FindAllMarkers() can finds all differentially expressed genes for each of the patients in this dataset. Some constraints can be used to filter these genes, as min.pct that test for genes that are very infrequently expressed, which has as default value 0.1. The results are joined and the gene markers are filtered only for cluster number 4.

```
patient_splitted <- SplitObject(covid_p3_p10, split.by = "patient")
p3_markers <- FindAllMarkers(object = patient_splitted$P3)
p10_markers <- FindAllMarkers(object = patient_splitted$P10)
p3_markers[["patient"]] = "P3"
p10_markers[["patient"]] = "P10"
p3_p10_markers <- rbind(p3_markers, p10_markers)
cluster_4_markers <- p3_p10_markers[which(p3_p10_markers["cluster"] == "4"),]
```

The next step is to group the expressed genes as “Not Significant,” “Significant,” “FoldChange,” and “Significant&FoldChange” depending on the values of p-value and fold change. A plot (Fig. 3.13) with the most significant differentially expressed genes for the patients P3 and P10 can be built to highlight them.

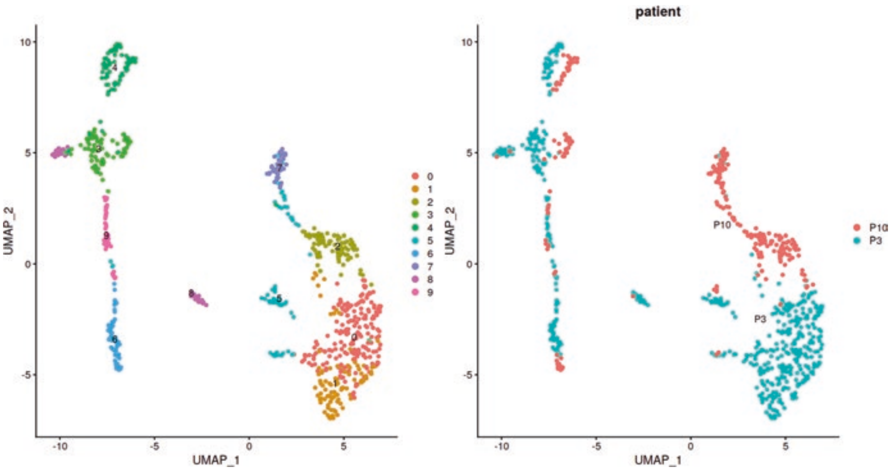


Fig. 3.12 Ten cell clusters belonging to the patients P3 and P10. Dimensionality reduction yields clusters of cells correlated by gene expression profile. Each cluster is labeled with a different color and is identified by a number that can be later annotated as a particular cell type based on the gene markers expressed in the cluster

```
# Preliminarly grouping all genes as "Not Significant"
cluster_4_markers["group"] <- "Not Significant"
# Change the grouping for the entries with significance but not a large enough Fold change
cluster_4_markers [which(cluster_4_markers["p_val_adj"] < 0.05 &
                     abs(cluster_4_markers["avg_log2FC"]) < 1 ),"group"] <- "Significant"
# Change the grouping for the entries a large enough Fold change but not a low enough p-value
cluster_4_markers [which(cluster_4_markers["p_val_adj"] > 0.05 &
                     abs(cluster_4_markers["avg_log2FC"]) > 1 ),"group"] <- "FoldChange"
# Change the grouping for the entries with both significance and large enough fold change
cluster_4_markers[which(cluster_4_markers["p_val_adj"] < 0.05 &
                    abs(cluster_4_markers["avg_log2FC"]) > 1 ),"group"] <- "Significant&FoldChange"
# Find and label the top peaks
top_peaks <- cluster_4_markers[which(cluster_4_markers["group"] == "Significant&FoldChange",
                               order(cluster_4_markers["p_val_adj"])),][1:10,]
p3_p10_plot <- ggplot(na.omit(cluster_4_markers)) +
  geom_point(aes(x = avg_log2FC, y = -log10(p_val_adj), colour = group, shape = patient), size = 5) +
  geom_text_repel(data=top_peaks[1:7,],aes(x = avg_log2FC, y = -log10(p_val_adj),label = gene))+
  scale_color_brewer(palette = "PuRd") +
  ggtitle("Most significant expressed genes in cluster 4 for patients P3 and P10") +
  xlab("log2 fold change") +
  ylab("-log10 adjusted p-value") +
  theme_minimal() +
  theme(legend.position = "bottom",
        legend.title = element_blank(),
        plot.title = element_text(size = rel(1), hjust = 0.5),
        axis.title = element_text(size = rel(1)))
p3_p10_plot
```

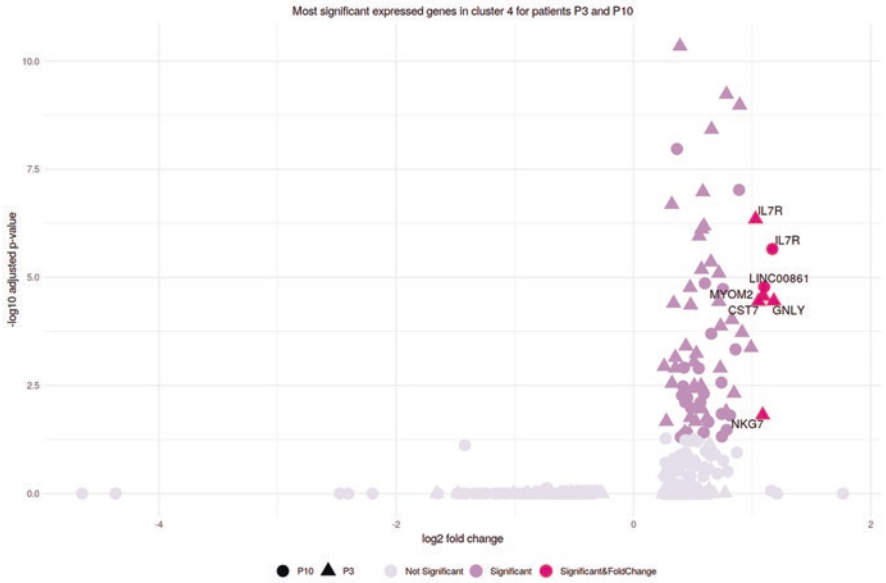


Fig. 3.13 Differentially expressed genes for patients 3 and 10. Each cluster of cell is tested against all remaining clusters. The most significant down- and upregulated genes are highlighted. Patient 3 is shown in the left and patient 10 in the right

The differentially expressed genes depicted in Fig. 3.13 reveal that six genes meet both statistical and fold change criteria. The IL7 receptor (IL7R) appears upregulated in both patients, while GNLY, MYOM2, CST7, and NKG7 are upregulated only in patient 3. The LincRNA 00861, a non-coding RNA, is upregulated only in patient 10, who had a milder infection. All of these genes are usually expressed in the cytotoxic CD8 lymphocytes, but patient 10, who evolved a strong inflammatory response, reveals a different gene response that is not associated with the LincRNA but strongly associated with genes involved in cytotoxicity (NKG7 and GNLY).

Single-cell computational analysis can consume vast computational resources. This case study uses only part of the original data to make it reproducible in a regular desktop or notebook computer. All these codes are available for download with the environment set-up instructions at https://github.com/waldeyr/single_cell_analysis.

Acknowledgments This work was supported by grants from CNPq, FAP-DF and CAPES/BioComp. FAH is supported by CAPES doctoral fellowship.

References

10x Genomics Inc (2020) Explore cellular diversity at scale. Product Sheet | Single Cell Gene Expression v3.1 with Feature Barcode technology. Pleasanton

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amarasinghe SL, Su S, Dong X et al (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21. https://doi.org/10.1186/s13059-020-1935-5

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11. https://doi.org/10.1186/gb-2010-11-10-r106

Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. Genome Res 22:2008–2017. https://doi.org/10.1101/gr.133744.111

Azizi E, Carr AJ, Plitas G et al (2018) Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 174:1293–1308.e36. https://doi.org/10.1016/j.cell.2018.05.060

Baran Y, Bercovich A, Sebe-Pedros A et al (2019) MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol 20. https://doi.org/10.1186/s13059-019-1812-2

Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. https://doi.org/10.1038/nature07517

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chi HW, Yang YS, Shang ST et al (2011) Candida albicans versus non-albicans bloodstream infections: the comparison of risk factors and outcome. J Microbiol Immunol Infect 44:369–375. https://doi.org/10.1016/j.jmii.2010.08.010

Deamer D, Akeson M, Branton D (2016) Three decades of nanopore sequencing. Nat Biotechnol 34:518–524. https://doi.org/10.1038/nbt.3423

Ding J, Adiconis X, Simmons SK et al (2019) Systematic comparative analysis of single cell RNA-sequencing methods. bioRxiv 632216. https://doi.org/10.1101/632216

Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635

Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(80):133–138. https://doi.org/10.1126/science.1162986

Gao S (2018) Data analysis in single-cell transcriptome sequencing. In: Methods in molecular biology. Humana Press, pp 311–326

Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. https://doi.org/10.1038/nbt.1883

Hagemann-Jensen M, Ziegenhain C, Chen P et al (2020) Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat Biotechnol 38:708–714. https://doi.org/10.1038/s41587-020-0497-0

Hao Y, Hao S, Andersen-Nissen E et al (2020) Integrated analysis of multimodal single-cell data. bioRxiv:2020.10.12.335331. https://doi.org/10.1101/2020.10.12.335331

Hardcastle TJ, Kelly KA (2010) BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinf 11. https://doi.org/10.1186/1471-2105-11-422

Hashimshony T, Senderovich N, Avital G et al (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol 17. https://doi.org/10.1186/s13059-016-0938-8

Hoffmann S, Otto C, Kurtz S et al (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5. https://doi.org/10.1371/journal.pcbi.1000502

Hoffmann S, Otto C, Doose G et al (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biol 15. https://doi.org/10.1186/gb-2014-15-2-r34

Hu Z, Lyu T, Yan C et al (2020) Identification of alternatively spliced gene isoforms and novel non-coding RNAs by single-molecule long-read sequencing in Camellia. RNA Biol 17:966–976. https://doi.org/10.1080/15476286.2020.1738703

Hünniger K, Lehnert T, Bieber K et al (2014) A virtual infection model quantifies innate effector mechanisms and Candida albicans immune escape in human blood. PLoS Comput Biol 10. https://doi.org/10.1371/journal.pcbi.1003479

Ip CLC, Loose M, Tyson JR et al (2015) MinION Analysis and Reference Consortium: phase 1 data release and analysis. F1000 Res 4. https://doi.org/10.12688/f1000research.7201.1

Islam S, Zeisel A, Joost S et al (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 11:163–166. https://doi.org/10.1038/nmeth.2772

Kanzi AM, San JE, Chimukangara B et al (2020) Next generation sequencing and bioinformatics analysis of family genetic inheritance. Front Genet 11. https://doi.org/10.3389/fgene.2020.544162

Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7:1009–1015. https://doi.org/10.1038/nmeth.1528

Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. https://doi.org/10.1186/gb-2013-14-4-r36

Klein AM, Mazutis L, Akartuna I et al (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

Köster J, Rahmann S (2012) Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kovaka S, Zimin AV, Pertea GM et al (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol 20. https://doi.org/10.1186/s13059-019-1910-1

Kuhn RM, Haussler D, James Kent W (2013) The UCSC genome browser and associated tools. Brief Bioinform 14:144–161. https://doi.org/10.1093/bib/bbs038

Kvam VM, Liu P, Yaqing S (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot 99:248–256. https://doi.org/10.3732/ajb.1100340

Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10. https://doi.org/10.1186/gb-2009-10-3-r25

Leśniewska A, Okoniewski MJ (2011) rnaSeqMap: a bioconductor package for RNA sequencing data exploration. BMC Bioinf 12. https://doi.org/10.1186/1471-2105-12-200

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li J, Witten DM, Johnstone IM, Tibshirani R (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics 13:523–538. https://doi.org/10.1093/biostatistics/kxr031

Li F, Luo M, Zhou W et al (2020) Single cell RNA and immune repertoire profiling of COVID-19 patients reveal novel neutralizing antibody. Protein Cell. https://doi.org/10.1007/s13238-020-00807-6

Logsdon GA, Vollger MR, Eichler EE (2020) Long-read human genome sequencing and its applications. Nat Rev Genet 21:597–614. https://doi.org/10.1038/s41576-020-0236-x

Lu H, Giordano F, Ning Z (2016) Oxford nanopore MinION sequencing and genome assembly. Genomics Proteomics Bioinf 14:265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Luecken MD, Theis FJ (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 15. https://doi.org/10.15252/msb.20188746

Macosko EZ, Basu A, Satija R et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Magen A, Nie J, Ciucci T et al (2019) Single-cell profiling defines transcriptomic signatures specific to tumor-reactive versus virus-responsive CD4+ T cells. Cell Rep 29:3019–3032.e6. https://doi.org/10.1016/j.celrep.2019.10.131

Maranhão AQ, Silva HM, da Silva WMC et al (2020) Discovering selected antibodies from deep-sequenced phage-display antibody library using ATTILA. Bioinf Biol Insights 14. https://doi.org/10.1177/1177932220915240

Marr KA, Patterson T, Denning D (2002) Aspergillosis pathogenesis, clinical manifestations, and therapy. Infect Dis Clin N Am 16:875–894. https://doi.org/10.1016/S0891-5520(02)00035-1

Marsh M, Tu O, Dolnik V et al (1997) High-throughput DNA sequencing on a capillary array electrophoresis system. J Capillary Electrophor 4:83–89

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10. https://doi.org/10.14806/ej.17.1.200

Martínez-Álvarez JA, Pérez-García LA, Flores-Carreón A, Mora-Montes HM (2014) The immune response against Candida spp. and Sporothrix schenckii. Rev Iberoam Micol 31:62–66. https://doi.org/10.1016/j.riam.2013.09.015

Metzker ML (2010) Sequencing technologies the next generation. Nat Rev Genet 11:31–46

Miceli MH, Díaz JA, Lee SA (2011) Emerging opportunistic yeast infections. Lancet Infect Dis 11:142–151. https://doi.org/10.1016/S1473-3099(10)70218-8

Miramón P, Kasper L, Hube B (2013) Thriving within the host: Candida spp. interactions with phagocytic cells. Med Microbiol Immunol 202:183–195. https://doi.org/10.1007/s00430-013-0288-z

Nakano K, Shiroma A, Shimoji M et al (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell 30:149–161. https://doi.org/10.1007/s13577-017-0168-8

Nattestad M, Goodwin S, Ng K et al (2018) Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Res 28:1126–1135. https://doi.org/10.1101/gr.231100.117

Parekh S, Ziegenhain C, Vieth B et al (2018) zUMIs – a fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience 7. https://doi.org/10.1093/gigascience/giy059

Park JH, Lee HK (2020) Re-analysis of single cell transcriptome reveals that the NR3C1-CXCL8-neutrophil axis determines the severity of COVID-19. Front Immunol 11. https://doi.org/10.3389/fimmu.2020.02145

Pereira R, Oliveira J, Sousa M (2020) Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J Clin Med 9:132. https://doi.org/10.3390/jcm9010132

Picelli S, Faridani OR, Björklund ÅK et al (2014) Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 9:171–181. https://doi.org/10.1038/nprot.2014.006

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramsköld D, Luo S, Wang YC et al (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30:777–782. https://doi.org/10.1038/nbt.2282

Rapaport F, Khanin R, Liang Y et al (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 14. https://doi.org/10.1186/gb-2013-14-9-r95

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinf 13:278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Richardson M, Lass-Flörl C (2008) Changing epidemiology of systemic fungal infections. Clin Microbiol Infect 14:5–24. https://doi.org/10.1111/j.1469-0691.2008.01978.x

Robertson G, Schein J, Chiu R et al (2010) De novo assembly and analysis of RNA-seq data. Nat Methods 7:909–912. https://doi.org/10.1038/nmeth.1517

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. https://doi.org/10.1093/bioinformatics/btp616

Robles JA, Qureshi SE, Stephen SJ et al (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13. https://doi.org/10.1186/1471-2164-13-484

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864. https://doi.org/10.1093/bioinformatics/btr026

Sedlazeck FJ, Rescheneder P, Von Haeseler A (2013) NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics 29:2790–2791. https://doi.org/10.1093/bioinformatics/btt468

Shendure J (2008) The beginning of the end for microarrays? Nat Methods 5:585–587. https://doi.org/10.1038/nmeth0708-585

Shigemura K, Osawa K, Jikimoto T et al (2014) Comparison of the clinical risk factors between Candida albicans and Candida non-albicans species for bloodstream infection. J Antibiot (Tokyo) 67:311–314. https://doi.org/10.1038/ja.2013.141

Simpson JT, Wong K, Jackman SD et al (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123. https://doi.org/10.1101/gr.089532.108

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Smith LM, Sanders JZ, Kaiser RJ et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679. https://doi.org/10.1038/321674a0

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3. https://doi.org/10.2202/1544-6115.1027

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinf 14. https://doi.org/10.1186/1471-2105-14-91

Soon WW, Hariharan M, Snyder MP (2013) High-throughput sequencing for biology and medicine. Mol Syst Biol 9. https://doi.org/10.1038/msb.2012.61

Tang S, Riva A (2013) PASTA: Splice junction identification from RNA-Sequencing data. BMC Bioinf 14. https://doi.org/10.1186/1471-2105-14-116

Tang F, Barbacioru C, Wang Y et al (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382. https://doi.org/10.1038/nmeth.1315

Tierney L, Linde J, Müller S et al (2012) An interspecies regulatory network inferred from simultaneous RNA-seq of Candida albicans invading innate immune cells. Front Microbiol 3. https://doi.org/10.3389/fmicb.2012.00085

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515. https://doi.org/10.1038/nbt.1621

Trapnell C, Hendrickson DG, Sauvageau M et al (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31:46–53. https://doi.org/10.1038/nbt.2450

Van Verk MC, Hickman R, Pieterse CMJ, Van Wees SCM (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci 18:175–179. https://doi.org/10.1016/j.tplants.2013.02.001

Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 131:281–285. https://doi.org/10.1007/s12064-012-0162-3

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. https://doi.org/10.1038/nrg2484

Wang B, Kumar V, Olson A, Ware D (2019) Reviving the transcriptome studies: an insight into the emergence of single-molecule transcriptome sequencing. Front Genet 10. https://doi.org/10.3389/fgene.2019.00384

Wang XJ, Jiao Y, Ma S et al (2020) Whole-genome sequencing: an effective strategy for insertion information analysis of foreign genes in transgenic plants. Front Plant Sci 11. https://doi.org/10.3389/fpls.2020.573871

Weirather JL, Afshar PT, Clark TA et al (2015) Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic Acids Res:43. https://doi.org/10.1093/nar/gkv562

Wenger AM, Peluso P, Rowell WJ et al (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wercelens P, da Silva W, Hondo F et al (2019) Bioinformatics workflows with NoSQL database in cloud computing. Evol Bioinforma 15. https://doi.org/10.1177/1176934319889974

Yu G, Wang LG, Han Y, He QY (2012) ClusterProfiler: an R package for comparing biological themes among gene clusters. Omi A J Integr Biol 16:284–287. https://doi.org/10.1089/omi.2011.0118

Zhang J, Su L, Wang Y, Deng S (2020a, 2020) Improved high-throughput sequencing of the human oral microbiome: from illumina to PacBio. Can J Infect Dis Med Microbiol. https://doi.org/10.1155/2020/6678872

Zhang YZ, Akdemir A, Tremmel G et al (2020b) Nanopore basecalling from a perspective of instance segmentation. BMC Bioinf 21. https://doi.org/10.1186/s12859-020-3459-0

Zhao L, Zhang H, Kohnen MV et al (2019) Analysis of transcriptome and epitranscriptome in plants using pacbio iso-seq and nanopore-based direct RNA sequencing. Front Genet:10. https://doi.org/10.3389/fgene.2019.00253

Zheng GXY, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8. https://doi.org/10.1038/ncomms14049

Zimin AV, Marçais G, Puiu D et al (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677. https://doi.org/10.1093/bioinformatics/btt476

Chapter 4
Transcriptomics of Neonatal and Infant Human Thymus

Carlos Alberto Moreira-Filho, Silvia Yumi Bando, Fernanda Bernardi Bertonha, and Magda Carneiro-Sampaio

4.1 Introduction

The thymus is a primary lymphoid organ where immunocompetent and self-tolerant T cells are produced (Kondo et al. 2019; Miller 2020). Anatomic descriptions of the thymus date back to Rufus of Ephesus and Galen, in 1st–2nd c A.D (Laios 2018). Yet, it was only in the 1970s and 1980s – with the discovery of the basic cellular and molecular mechanisms for generating B and T cell antigen receptor diversity, and for the elimination of self-reactive T cells in the thymic stroma – that our current understanding of thymic functions became established (reviewed in Geenen 2021). Presently, remarkable progresses in genomics and systems biology – mainly on single-cell RNA sequencing (scRNA-seq) methodologies and on computational analysis of genomic and proteomic big data – have allowed the construction of detailed atlases of thymic organogenesis and development (Kernfeld et al. 2018; Ohigashi et al. 2019; Park et al. 2020; Hao et al. 2021). By using scRNA-seq and spatial cell localization techniques, Park et al. (2020) constructed a temporal scenario of the human thymus cell state dynamic changes, spanning from organ development to pediatric and adult life. The human thymus involutes very early in life (Steinmann 1986; Rezzani et al. 2014) and its functioning suffers the influence of sex hormones (Dragin et al. 2016; Berrih-Aknin et al. 2018; Merrheim et al. 2020). Therefore, the thymus early programming – sexual dimorphism, dynamics of thymocyte populations, and thymic microenvironment change – influences immune activity throughout life. Here, we used transcriptome analysis to address two key issues in early thymus development: the thymic sexual dimorphism during the first

C. A. Moreira-Filho (✉) · S. Y. Bando · F. B. Bertonha · M. Carneiro-Sampaio
Departamento de Pediatria, Faculdade de Medicina da Universidade de São Paulo,
São Paulo, SP, Brazil
e-mail: carlos.moreira@hc.fm.usp.br; silvia.bando@hc.fm.usp.br;
fernanda.bernardi@fm.usp.br; magdacs@usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_4

6 months of age (i.e., during minipuberty), and the genomic mechanisms governing cellular and molecular processes involved in the functioning of the neonate thymus and in the onset of thymic decline.

4.2 Methodology

Fresh thymic explants (corticomedullar sections) from karyotypically normal neonates and infants who underwent cardiac surgery were collected at surgery room and immediately preserved for total RNA extraction. The resected thymic specimens were preserved in formalin and paraffin-embedded for histological analyses. DNA microarray technology was employed to obtain mRNA and miRNA expression profiles. Transcriptomics was performed in whole thymic tissue to avoid gene expression artifacts caused by mechanic/enzymatic tissue dissociation. Whole thymic tissue transcriptome datasets were interpreted through modular repertoire identification, a community detection network analysis (Barabási and Oltvai 2004; Zhu et al. 2007; Barabási et al. 2011; Chaussabel and Baldwin 2014; Gaiteri et al. 2014; van Dam et al. 2018) used, for instance, in the investigation of immune responses in vivo following the administration of vaccines (Nakaya et al. 2011; Obermoser et al. 2013) and that has proved to be a suitable strategy to circumvent tissue microdissection and cell separation (Moreira-Filho et al. 2015, 2016; Bando et al. 2021).

For the investigation of thymic sexual dimorphism, we adopted a network-based approach for gene co-expression (GCN) analysis that allows the identification of modular transcriptional repertoires (communities) and the interactions between all the system's constituents through community detection (Chaussabel and Baldwin 2014; Moreira-Filho et al. 2016). MiRNA-target analysis was used to investigate how the abundantly expressed thymic miRNAs modulate the expression of highly connected genes (hubs) in the GCNs. In the study of neonatal and infant thymus we employed a weighted gene co-expression network analysis (WGCNA) (Langfelder and Horwath 2008) for describing correlation patterns among genes across microarray datasets that allows: (i) the identification of transcriptional modules (van Dam et al. 2018) and their association with specific age groups; (ii) the identification of highly connected genes (hubs) and of significant genes (HGS genes) for the trait of interest (age). This analysis was complemented by an integrative mRNA–miRNA–transcription factor (TF) co-expression analysis encompassing mRNAs from hubs and HGS genes, the abundantly expressed miRNAs, and the TFs that covaried with hubs and/or HGS genes. The methodological framework employed these two studies is detailed in chapter 6. The workflow of experimental approaches and bioinformatic analyses is depicted in Fig. 4.1.

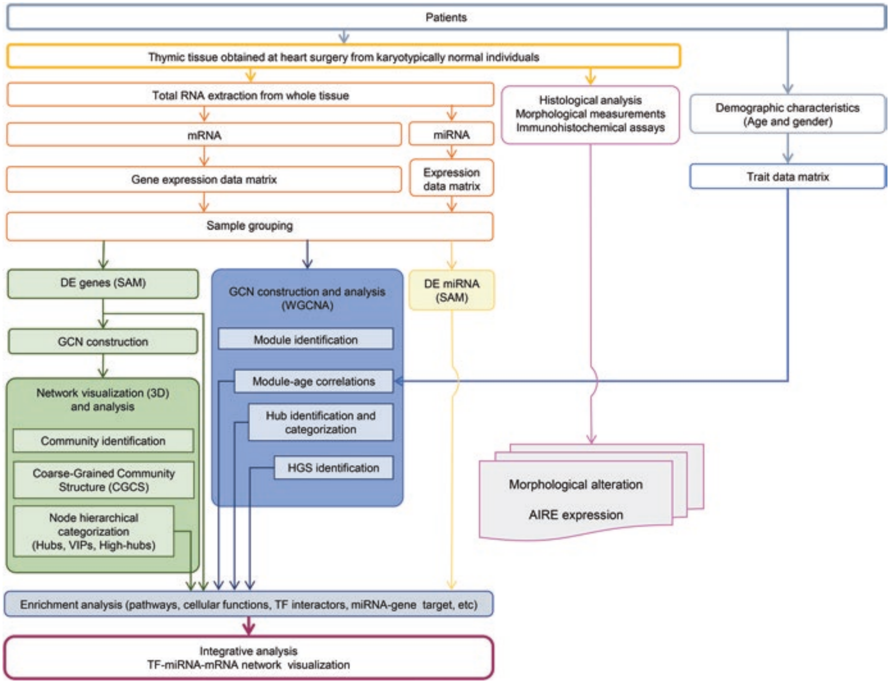

Patients
Thymic tissue obtained at heart surgery from karyotypically normal individuals
Total RNA extraction from whole tissue
Histological analysis
Morphological measurements
Immunohistochemical assays
Demographic characteristics
(Age and gender)
mRNA
miRNA
Gene expression data matrix
Expression data matrix
Trait data matrix
Sample grouping
DE genes (SAM)
GCN construction and analysis (WGCNA)
DE miRNA (SAM)
GCN construction
Module identification
Module-age correlations
Network visualization (3D) and analysis
Hub identification and categorization
Community identification
HGS identification
Coarse-Grained Community Structure (CGCS)
Morphological alteration
AIRE expression
Node hierarchical categorization
(Hubs, VIPs, High-hubs)
Enrichment analysis (pathways, cellular functions, TF interactors, miRNA-gene target, etc)
Integrative analysis
TF-miRNA-mRNA network visualization

Fig. 4.1 Systems biology workflow used for generating and integrating data on neonatal and infant thymus development

4.3 Human Thymus During Minipuberty

Sexual dimorphism in the immune system is well documented in humans – as well as in other mammals and birds – encompassing sex differences in responses to self and foreign antigens: women usually mount stronger immune responses to infections and vaccination but have higher susceptibility to autoimmune diseases than men (Klein and Flanagan 2016). Regarding autoimmune diseases, it is striking that in USA almost 80% of autoimmune patients are women (Billi et al. 2019). Autoimmunity results from a tolerance breakdown and essentially involves the thymus, the site of T cell selection (Cheng and Anderson 2018). T cell selection depends on the ectopic thymic transcription of thousands of genes coding for tissue-specific antigens, which is induced by the autoimmune regulator gene *AIRE* (Passos et al. 2018; Perniola 2018). Despite our incomplete knowledge on the biological processes responsible for autoimmunity, it would be reasonable to assume that sex hormones impact the genomic mechanisms governing *AIRE* functions and T cell selection. An important experimental evidence regarding this assumption came from the work of Dumont-Lagacé et al. (2015), who showed in a murine model that sex hormones have pervasive effects on thymic epithelial cells (TEC) – antigen presenting cells that regulate T cell repertoire and tolerance – and that androgens have a greater impact on TEC transcriptome than estrogens. Interestingly, the authors observed that sex steroids repressed the expression of tissue-restricted antigens but did not alter the expression of *Aire*. Just after this work, Dragin et al. (2016) demonstrated that estrogen mediates the downregulation of *AIRE* in human pubescent and adult thymic tissues, thus indicating that the reduced expression of AIRE protein in women may be related to autoimmunity susceptibility. However, this study did not cover infants along the first 6 months of age, i.e., during minipuberty (Kuiri-Hänninen et al. 2014; Becker and Hesse 2020), a period when sex hormones conceivably act on thymic tissue.

To further investigate the presumptive sexual dimorphism induced by minipuberty on infant thymus, we performed comparative genomic and immunohistochemical studies on thymic surgical explants obtained from karyotypically normal male (M) and female (F) infants during minipuberty, here termed MM and MF groups, and from karyotypically normal M and F non-puberty (N) infants aged 7–31 months, the NM and NF groups. Analyses included gene co-expression networks (GCN) for differentially expressed genes, miRNA-target analyses, *AIRE*-centered gene–gene interaction networks encompassing the genes coding for AIRE interactors, quantitative RT-qPCR and immunohistochemical measurements of AIRE expression, and comparative thymic histomorphometry. GCN analysis was performed for the identification of high-hierarchical genes, modular transcriptional repertoires (communities), and the interactions between all the system's constituents through community detection (Moreira-Filho et al. 2016). MiRNA-target analysis was used to investigate how the abundantly expressed thymic miRNAs modulate the expression of highly connected genes (hubs) in the GCNs. This study is fully described in Moreira-Filho et al. (2018).

4.4 Thymic Gene Expression in Minipuberty and Non-puberty Infants: mRNA, miRNA, and *AIRE* Interactors

4.4.1 Sample Grouping

Thymic samples from 17 patients aged up to 6 months were classified as minipuberty (M) – ten males and seven females, here termed MM and MF, respectively – and 17 samples from patients aged 7–17 months were classified as non-puberty (N), being nine males and eight females, here termed NM and NF, respectively.

4.4.2 Global Gene Expression and miRNA Expression in Minipuberty and Non-puberty Groups

DNA microarray technology was used to obtain mRNA and miRNA expression profiles in minipuberty (M) and non-puberty (N) groups. The mRNA expression data was analyzed by SAM test (Tusher et al. 2001) for determining the differentially expressed (DE) genes. In the MM *vs.* MF group comparison 494 DE genes were identified, all being hyper-expressed in the MM group. No DE genes were found in the NM *vs.* NF group comparison. Hence, we conducted DE network analysis only for MM and MF groups. Statistical analysis (t-test) for miRNA expression data revealed 16 abundantly expressed miRNAs in the M group, being all hyper-expressed in the MF group (female infants). In the non-puberty groups, 20 abundantly expressed miRNAs (15 of which were also abundantly expressed in the minipuberty groups) and all hyper-expressed in the NM group (male infants). The abundantly expressed miRNAs for minipuberty and non-puberty groups were selected after analyzing miRNA expression value distribution through a scatter dot plot and adopting abundant expression cut-off values for MM and MF groups, and for NM and NF groups, respectively. The fold change values were calculated as the ratio of the average expression value of each abundantly expressed miRNA to the average expression value of the non-abundantly expressed miRNAs for each group.

4.4.3 Gene Co-expression Network Construction and Analysis

Differentially expressed (DE) GO annotated gene co-expression networks were constructed for MM and MF groups based on gene–gene Pearson's correlation method and the Networks 3D software developed by Luciano Costa's Research Group, Institute of Physics at São Carlos, University of São Paulo (Bando et al. 2013). This package allowed the categorization of network nodes according to distinct hierarchical levels of gene–gene connections: hubs are highly connected nodes, VIPs have low node degree but connect only with hubs, and high-hubs have VIP status and high overall number of connections. We classified network nodes as VIPs, hubs, or high-hubs by obtaining the node degree, k_0, and the first level concentric node degree, k_1, which takes into account all node connections leaving from its immediate neighborhood, then projecting all node values in a k_0 vs k_1 graphic (see Chap. 6 and Bando et al. 2013).

The GCN topologies were then analyzed for identifying their community structure. Communities can encompass complex mechanisms that work together to maintain the cellular processes across different conditions (Barabasi et al. 2011; Gaiteri et al. 2014; van Dam et al. 2018). For example, community structure analysis of gene co-expression networks obtained from skeletal muscle cells' transcriptome revealed different biological pathways for Duchenne muscular dystrophy patients comparatively to normal individuals (Narayanan and Subramaniam 2013). The same modular approach has successfully been used for investigating immune response to infections and vaccines using whole blood transcriptome data sets (reviewed in Chaussabel and Baldwin 2014), and for characterizing thymic gene dysregulation in 21 trisomy patients (Moreira-Filho et al. 2016).

Community detection in complex networks is usually accomplished by discovering the network modular structure that optimizes the modularity measurement. Modularity considers the relationship between the number of links inside a community compared to connections between nodes in distinct communities (Newman and Girvan 2004; Newman 2010). A diverse range of optimization techniques exists to optimize the modularity. Here we applied the method proposed by Blondel et al. (2008), which attains good modularity values and at the same time presents excellent performance.

Most of the methods for community detection generate hierarchical structures. The Newman-Girvan method uses the edge betweenness centrality measurement as a criterion for removing edges and obtaining connected components that correspond to each network partition. This builds a tree of communities with branches occurring every time a component is divided into two. Agglomerative methods start from a set of communities, where each node corresponds to a different community, which are progressively merged according to a similarity criterion or to directly maximize the change of modularity (Clauset et al. 2004). In both cases, a dendrogram of the partition hierarchy is obtained. The optimal set of communities is then obtained by a cut for the highest value of modularity.

Figure 4.2a, b depicts the two minipuberty networks, MM-DE and MF-DE, their gene communities (modules), and the high-hierarchy genes for each network. Different node colors identify the distinct gene communities in each network. Modularity values and the number of communities in each network were quite close: 0.728 and 15 communities in the MM-DE and 0.649 and 16 communities in MF-DE (Moreira-Filho et al. 2018). Coarse-grained community structure (CGCS) was obtained for each DE network, disclosing the relationships between each community in the network (Fig. 4.2c, d) for MM-DE and MF-DE, respectively). Communities having the highest node strength (total probability for community's nodes to connect to distinct communities) hold the most significant functional interactions in the network (Chaussabel and Baldwin 2014).

The integrative network analyses between abundantly expressed miRNAs and target high-hierarchy genes (HH) from MM-DE and MF-DE networks appear in Fig. 4.2a, b. It is worth to note that all miRNAs interacting with HH genes in the MM-DE and MF-DE networks play important roles in the regulation of immune processes, and particularly in the thymic environment. Let-7 miRNAs regulate NKT cell differentiation (Pobezinsky et al. 2015). The cluster miR15/16 enhances the induction of regulatory T-cells by regulating the expression of Rictor and TOR

Fig. 4.2 (continued) High hierarchy genes are identified by their node border color: green for high-hubs, red for VIPs, and blue for hubs. Abundantly expressed miRNAs are depicted as vee nodes. Gray lines indicate gene–gene links, whereas miRNA-gene validated interactions are indicated by blue lines. The vees filled with red or green colors indicate, respectively, hyper- or hypo-expressed miRNAs. Gene communities in both network diagrams are distinguished by different node colors. In CGCS, the communities are identified by different colors and the edge width and intensity are proportional to the connection weight of edges linking distinct communities. In the networks, the node size is proportional to the number of gene–gene links. In CGCS diagrams, the node size is proportional to the number of nodes/genes in each community. In the MM-DE network, the communities harboring high hierarchy genes are identified by the following colors: A, blue; B, orange; D, red; F, brown; G, pink; and I, olive green. In the MF-DE network the communities and their respective colors are: A, blue; B, orange; C, green; D, red; and E, purple

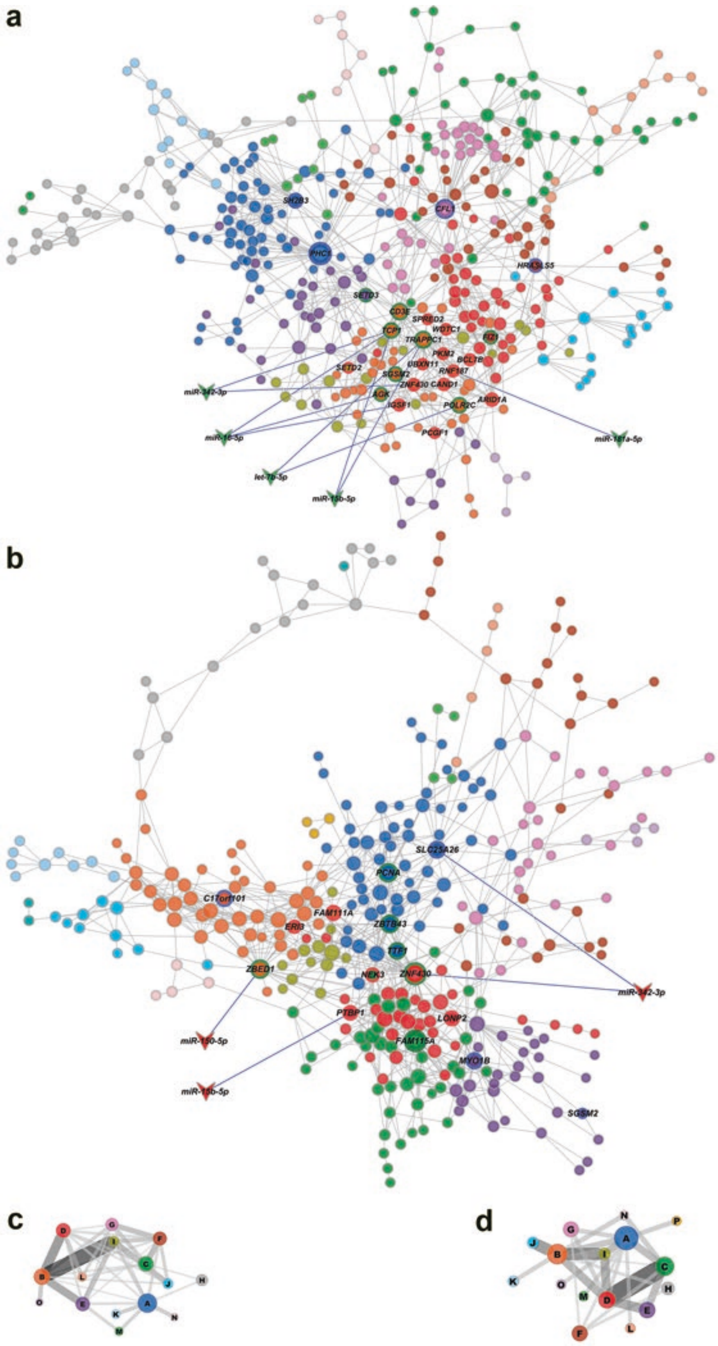


Fig. 4.2 DE networks with their respective gene communities (modules), miRNA–target interactions, and coarse-grained community structure (CGCS) diagrams. Network topology and community structure for minipuberty DE networks (**a** for MM and **b** for MF), and CGCSs for minipuberty DE networks (**c** for MM and **d** for MF) considering 15 and 16 communities per network, respectively.

(Singh et al. 2015). MiR-150 controls the Notch pathway and influences T-cell development and physiology (Ghisi et al. 2011). MiR-181 enhances cell proliferation in medullary thymic epithelial cells via regulating TGF-β signaling (Guo et al. 2016) and is involved in the positive and negative selection of T-cells (Fu et al. 2014). MiR-342-3p is a well-known regulator of the NF-κB pathway (Zhao and Zhang 2015), whose activation was shown to be necessary for the thymic expression of Aire in mice (Zhu et al. 2006; Haljasorg et al. 2015).

In the MM-DE network (Fig. 4.2a) community B harbors most of the HH genes (17 out of 24) and all the interactions between HH genes and abundantly expressed miRNAs. Moreover, all the HH genes in community B are VIPs (11 genes) or high-hubs (six genes), which means that these genes play relevant roles regarding the network functioning and robustness (van Dam et al. 2018). Indeed, VIPs connect different gene communities (Bando et al. 2013) and high-hubs are essential for the maintenance of network robustness (Azevedo and Moreira-Filho 2015). Network biology studies have shown that GCNs can be effectively used to associate highly connected genes (i.e., GCN hubs) with biological functions/processes in cells and tissues (Zhu et al. 2007; Gaiteri et al. 2014). Targeted hub attacks in protein–protein and gene–gene networks have been used to disclose relevant functional genes in health and disease (Gaiteri et al. 2014; Azevedo and Moreira-Filho 2015; Farooqui et al. 2018). Therefore, GCN hubs are relevant for both network topology and cell functioning.

Noteworthy, miRNA-target interactions involved only VIPs and high-hubs in MM-DE network. One of these high-hubs, TCP1, which codes for a molecular chaperone required for the transition of double negative to double positive T cells in the thymus (Cao et al. 2008), has interactions with three abundantly expressed miRNAs, all exerting known regulatory roles in the immune system. Functionally, most of the HH genes in MM-DE network are related to DNA and chromatin binding, DNA repair, histone modification, and ubiquitination (Moreira-Filho et al. 2018). CGCS analysis shows clearly that community B holds the highest connection weights, thus evidencing its importance for network functioning (Fig. 4.2c).

In the MF-DE network (Fig. 4.2b), the HH genes are quite evenly distributed among five gene communities: A (three high-hubs and one hub), B (two VIPs, one high-hub, and one hub), C (one high-hub and one VIP), D (two VIPs and one high-hub), and E (two hubs). Abundantly expressed miRNAs were found to interact with two high-hubs, one VIP, and one hub. The genes involved in these interactions were related to DNA binding (two genes), alternative mRNA splicing (one gene), and transmembrane (mitochondrial) transporter activity (one gene). The most represented molecular functions and biological processes among HH genes in MF-DE network are related to DNA binding, control of gene expression, and DNA repair and replication. CGCS analysis shows that the five gene communities harboring HH genes are also the ones presenting the highest connection weights (Fig. 4.2d).

GCN analyses (Fig. 4.2a, b) clearly show that abundantly expressed miRNAs interact almost exclusively with high-hubs and VIPs, i.e., with genes that are essential for network robustness (high-hubs) and for connecting gene communities

(VIPs). Altogether, these results indicate that testosterone and estradiol surges in minipuberty are related to significant changes in HH genes in MM and MF networks, respectively, and that these changes are under tight control by abundantly expressed miRNAs interacting with high-hubs and VIPs. In fact, relevant thymic functions, such as the induction of regulatory T cells, are regulated by abundantly expressed miRNAs (Singh et al. 2015). Noteworthy, all miRNAs interacting with HH genes in both networks play important roles in the regulation of immune processes, and particularly in the thymic environment.

4.4.4 AIRE *Interactors*

Since *AIRE* expression assessment by microarray analysis, RT-qPCR, and immunohistochemistry revealed no significant differences between male and female groups (Moreira-Filho et al. 2018), we constructed four other GCNs for investigating *AIRE* interactors' gene–gene expression relationships for minipuberty (MM and MF) and non-puberty groups (NM and NF). These *AIRE* interactors networks included *AIRE* and other 34 genes (34 genes in the minipuberty group and 33 genes in the non-puberty group), which code for proteins that are associated, directly or indirectly, with AIRE and exert impact on its functions (Abramson et al. 2010; Abramson and Goldfarb 2016). *AIRE* interactors were classified according to their molecular function and represented by different node colors in the networks (Fig. 4.3). Gene–gene expression relationships of *AIRE* interactors presenting a Pearson's correlation coefficient value ≥0.70 at least in one group across minipuberty and non-puberty samples were termed high interactors. We found 14 high-interactors distributed among minipuberty and non-puberty groups and, noteworthy, distinctive profiles of AIRE interactors' gene–gene relationships for each group (Fig. 4.3). The MM group encompassed more high interactors (seven) than the other three groups. These data suggest that sex hormones and genomic background exert their influence on AIRE interactors' gene–gene expression relationship during and after minipuberty.

Interestingly, neither the sex steroids surge during minipuberty nor the XY or XX background were found to promote any significant gender-related histomorphometric changes – average cortical thickness; average diameter of the medullary region; total area of the lobule; area of the medullary region; and medullary area/lobule area (%) – in neonatal and infant thymus (Moreira-Filho et al. 2018), thus corroborating previous data (Steinmann et al. 1985; Steinmann 1986).

In conclusion, these results suggest that genomic mechanisms and postnatal hormonal influences probably act synergistically in shaping thymic sexual dimorphism along the first 6 months of life, but this process does not involve changes in *AIRE* expression, although may involve differences – perhaps long-lasting differences – in the interactions of AIRE with its partners.

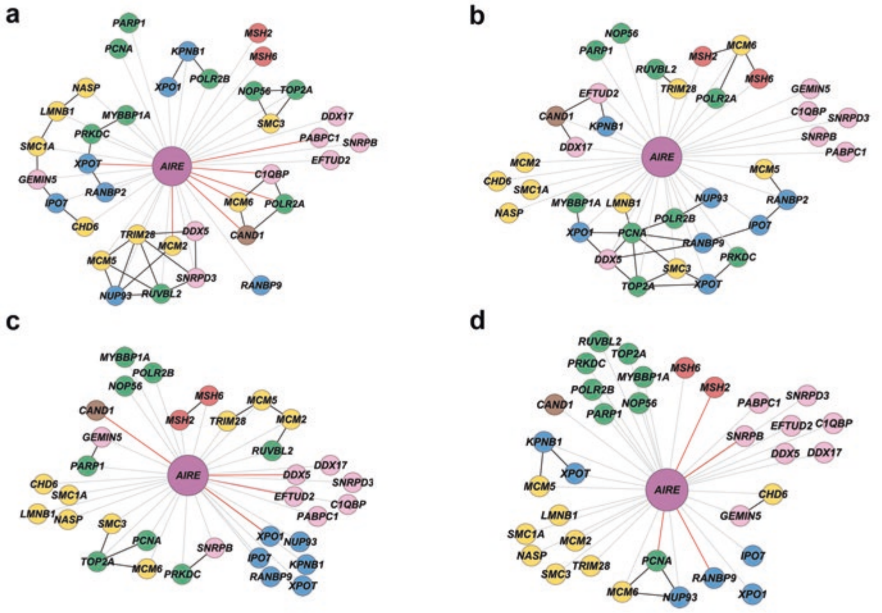


Fig. 4.3 AIRE interactors' gene–gene expression relationships. Gene–gene expression relationship networks for MM (**a**), MF (**b**), NM (**c**), and NF (**d**) groups. Nodes are colored according to their molecular function (GO): green for transcription, yellow for chromatin binding/structure, blue for nuclear transport, brown for ubiquitination, pink for pre-mRNA processing, red for DNA repair, and purple for *AIRE*. *AIRE*–gene expression correlation values <|0.70| are depicted with gray links; *AIRE*–gene expression correlation values ≥|0.70| are depicted with red links; gene–gene expression correlation values ≥|0.90| are depicted with black links

4.5 Age-Related Transcriptional Modules and TF–miRNA–mRNA Interactions in Neonatal and Infant Human Thymus

The human thymus grows only during the first year of life and its steady involution begins thereafter (Steinmann 1986; Thapa and Farber 2019). Moreover, the human thymus presents some functional peculiarities in the neonatal period and along the first 6 months of age, i.e., during minipuberty, when a transient surge in gonadal hormones takes place (Kuiri-Hänninen et al. 2014; Becker and Hesse 2020). In the neonatal thymus occurs a transient involution marked by severe depletion of double-positive (DP) thymocytes, which is later compensated by increased levels of primitive T-cell precursors. Concomitantly, there is a reinforcement of the subcapsular epithelial cell layer and an increase of the intralobular extracellular matrix network, leading to augmented thymic permeability and to the recirculation of primitive precursors and mature T-cells in the neonatal thymus (Varas et al. 2000). After the first year of life the total amount of lymphatic thymic tissue declines 5% per year until the 10th year and at progressively slower rates afterwards (Steinmann 1986). The histomorphological features of thymic postnatal growth and of infant and adult thymic aging (lymphatic tissue decline, lipomatous atrophy) were quite well studied (Steinmann et al. 1985; Steinmann 1986; Chinn et al. 2012; Gui et al. 2012; Cowan et al. 2020), but the genomic mechanisms underlying this process remain largely unknown.

4.5.1 Transcriptome Analysis

In order to further investigate the genomic mechanisms involved in thymic growth and early involution, we performed a comparative transcriptome analysis of whole thymic tissue from human neonates and from infants grouped according to sequential age intervals (6 months) up to the first 2½ years of life. Thymic tissue samples were obtained from 57 karyotypically normal patients who underwent cardiac surgery. For genomic analyses samples were classified according to patients' age in five sequential age groups: A, neonates up to 30 days; B, infants aged 31 days to 6 months; C, infants aged 7–12 months; D, infants aged 13–18 months; and E, patients aged 19–31 months. DNA microarray technology was employed to obtain mRNA and miRNA expression profiles.

Whole thymic tissue transcriptome datasets were interpreted through modular repertoire identification. Here we employed a weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath 2008) for describing correlation patterns among genes across microarray datasets that allows: (i) the identification of transcriptional modules (Chaussabel and Baldwin 2014) and their association with particular age groups; (ii) the identification of highly connected genes (hubs) and of significant genes (HGS genes) for the trait of interest (age). This analysis was complemented by an integrative mRNA–miRNA–transcription factor (TF) co-expression analysis encompassing mRNAs from hubs and HGS genes, the abundantly expressed miRNAs, and the TFs that covaried with hubs and/or HGS genes.

The mRNA expression data matrix was used for GCN construction employing the WGCNA package (Langfelder and Horvath 2008). After dynamic tree cut, the hierarchical clustering dendrogram identified 15 distinct gene modules (Fig. 4.4a, b) containing from 85 (midnight blue module) to 403 genes (turquoise module). Genes not clustered in any module were grouped in the grey module. Subsequently, each age group was correlated with all the co-expression modules. This module-trait correlation analysis revealed three modules – tan, green yellow, and brown – that were significantly ($p < 0.05$) associated with at least one age group (Fig. 4.4c). The green yellow module was positively correlated with group E (MS = 0.41, $p = 0.003$); the brown module was negatively correlated with group E (MS = −0.34, $p = 0.02$); while the tan module was negatively correlated with group A (MS = −0.31, $p = 0.03$), and it was positively and significantly correlated with group E (MS = 0.30, $p = 0.03$). None of the modules were significantly correlated with gender or with the age groups ranging from 31 days to 18 months.

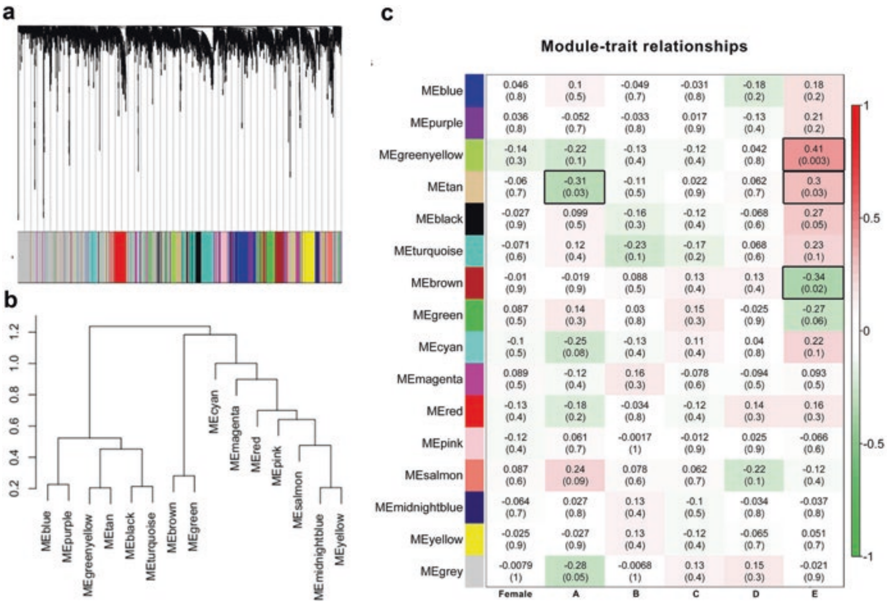


Fig. 4.4 WGCNA analysis. Gene dendrogram and gene clustering analysis for module identification (**a**). Hierarchical clustering dendrogram of the module-eigengenes (**b**). Module-trait relationships (**c**). The modules' names correspond to their colors (rows). Each column indicates a specific trait. The numbers inside each colored box are the module significance (MS) correlation values for gender and age groups, with p-value between parentheses. The more intense the box color, the more negatively (green) or positively (red) correlated is the module with the trait (MS values are indicated at the right color bar). Black-border boxes highlight the significant module-trait relationships

Hub genes identification in each significant module was accomplished through intramodular connectivity measures, i.e., the network nodes presenting high *k*Within values. A total of 34 hubs were found and assessed by an enrichment analysis. The Enrichr online web-based tool (Chen et al. 2013; Kuleshov et al. 2016) was used to identify significantly over-represented terms on GO Biological Process, Transcription Factor–PPIs Database, and miRTarBase. The TF–miRNA–mRNA regulatory network was then visualized by using the Cytoscape software, version 3.8.2 (Shannon et al. 2003). The hubs were mostly related to cellular/metabolic processes or related to T-cell development. The tan module (negatively associated with group A and positively associated with group E) encompasses a total of nine hubs. Two of them – *CAND1* and *ZNF675* – are related to medullary thymic epithelial cells (mTECs). The green yellow module (positively associated with group E) has a total of seven hubs. Three of them – *SNX17*, *MTMR4*, and *NKIRAS2* – are related to T-cell receptor (TCR) and thymic stromal functions. The brown module (negatively associated with group E) harbors a total of 18 hubs. Six of them – *CHMP5*, *PIK3CA*, *ARL8B*, *RNF138*, *NMI*, and *NRIP1* – are involved in T-cell development and antigen presentation-related functions.

The three age-related transcriptional modules here identified are correlated with two distinct and characteristic time intervals in human thymic evolution during the first 2½ years of life: the neonatal period (age group A) and the fourth/fifth semester period (age group E). In the neonatal period a transient thymic involution takes place, marked by severe cortical DP thymocyte depletion, and pronounced changes in the extracellular matrix network (Varas et al. 2000), whereas in the fourth semester of life the thymic involution becomes histomorphologically patent through the initial decline of the total amount of lymphatic tissue (Steinmann 1986). Interestingly, no transcriptional module was correlated with any interval in the 31 days to 18 months period, along which the thymus reaches its maximal growth (Steinmann 1986). Our TMA-IHC data on thymic cell subpopulations reflects this scenario, showing a continuous and moderate increase of thymocyte and B-cell numbers. These findings indicate that a genomic mechanism may act, synergistically with physiological and environmental stimuli (Moreira-Filho et al. 2018; Gui et al. 2012), on early thymic evolution/involution programming.

Additionally, we were able to identify the high gene significance (HGS) genes of the three modules significantly associated with age groups A (tan) and/or E (tan, green yellow, and brown). This categorization was accomplished according to module membership (MM) and gene significance (GS) values for groups A or E. Among the 50 HGS genes, a set of 37 were found to be DE: 19 genes were hyper-expressed in group A and 18 in group E. Moreover, 34 of these DE genes significantly varied their expression across all age groups. Among the hyper-expressed genes in group A, six are related to T-cell development and 13 to other cellular and metabolic processes. The hypo-expressed genes encompassed one gene related to T-cell development, six related to signaling pathways, and 11 related to other cellular and metabolic processes. It is interesting to mention that three of the hyper-expressed genes in the group A presented high fold-change values (>2.0): *CD5* and *CAND1*, in the tan module, and *SCML4*, in the green yellow module.

The integrative analysis of hubs and HGS genes, abundantly expressed miRNAs, and transcription factors (TFs) was accomplished by Pearson's correlation. For network construction we first obtained a gene expression data matrix for the above-mentioned genes, miRNA, and TFs. Subsequently, (i) for miRNA expression data we identified those abundantly expressed in at least one age group, and (ii) for TF expression data we used Enrichr online web-based tool (Chen et al. 2013; Kuleshov et al. 2016) to search TFs that have protein–protein interactions with hubs and HGS genes.

An integrative TF–miRNA–mRNA co-expression network of the abundantly expressed miRNAs covarying with hubs, HGS genes, and TFs was subsequently constructed (Fig. 4.5). It shows that most of the hub–hub or HGS–HGS gene links have positive correlations, while many hub–HGS gene links present negative correlations. Moreover, there are more hub–hub links than hub–HGS genes or HGS–HGS links. This result indicates that hubs are related to network module robustness and the HGS genes – which are differentially expressed genes – are either bridges between modules or border genes.

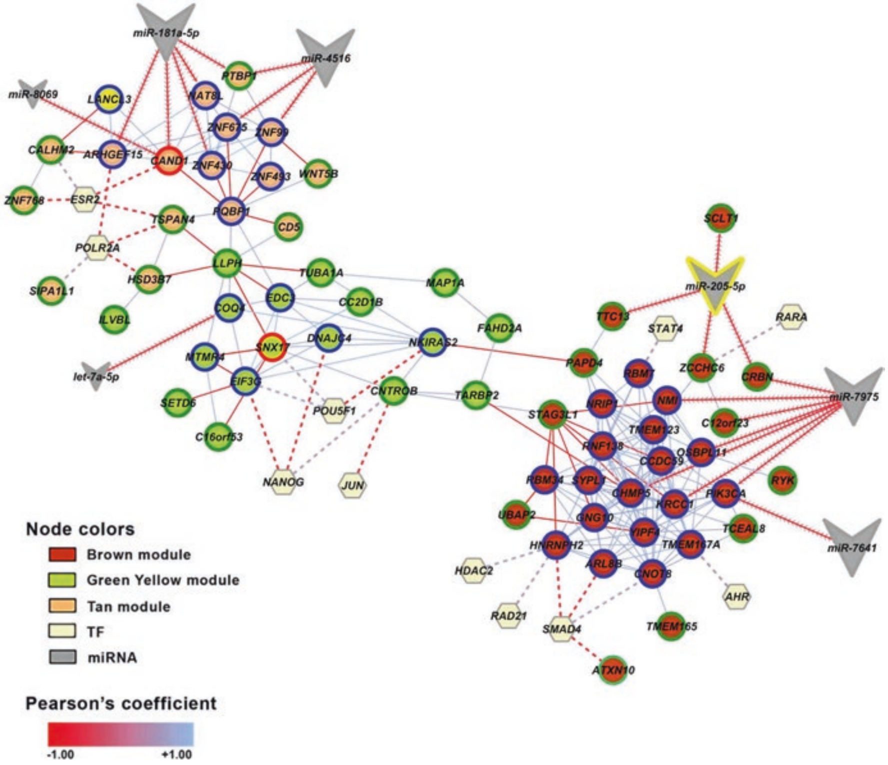


Fig. 4.5 Integrative TF–miRNA–mRNA co-expression subnetwork of the abundantly expressed miRNAs covarying with hubs, HGS genes, and transcription factors (TFs). Only co-expression covariance values of ≥|0.70| between gene–gene (solid lines), ≤ −0.50 between gene-miRNA (arrowed lines), and ≥|0.50| gene-TFs (dashed lines) were considered. Abundantly expressed miRNAs are depicted by gray vees; abundantly expressed and DE miRNAs are highlighted with a yellow border; HGS genes are depicted by green border nodes; hubs are depicted by blue border nodes; the two HGS genes that are also a hub gene are depicted by red border nodes; TFs are depicted by light yellow hexagons; positive and negative co-expression interactions are depicted by blue and red links, respectively

This integrative analysis clearly showed that the three age-related modules and their respective hubs are regulated by different and quite specific sets of abundantly expressed miRNAs and TF–hub interactions. The same situation prevails for the HGS genes, though it should be noted that just three TFs and several abundant miRNAs interact with the hyper-expressed genes in the age group A, whereas six different TFs but no abundantly expressed miRNA interact with the hypo-expressed genes in this group. The validated TF–miRNA interactions occurred more frequently with miR-150-5p, miR-181a-5p, and miR-205-5p.

Altogether, our results (Bertonha et al. 2020) show a genomic mechanism differentially governing the cellular and molecular processes involved in the functioning of the neonate thymus and, later, in the beginning of thymic decline. Along the first 2 years of age, this mechanism is tightly regulated by the differential expression of HGS genes and by TF-miRNA-hub/HGS interactions.

Acknowledgments This work was funded by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) research grants 2015/22308-2 (CAM-F) and 2014/50489-9 (MC-S); and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grant 3068936/2018-5 (CAM-F).

References

Abramson J, Goldfarb Y (2016) AIRE: From promiscuous molecular partnerships to promiscuous gene expression. Eur J Immunol 46:22–33

Abramson J, Giraud M, Benoist C et al (2010) AIRE's partners in the molecular control of immunological tolerance. Cell 140:123–135

Azevedo H, Moreira-Filho CA (2015) Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma. Sci Rep 5:16830

Bando SY, Silva FN, Costa L et al (2013) Complex network analysis of CA3 transcriptome reveals pathogenic and compensatory pathways in refractory temporal lobe epilepsy. PLoS One 8:e79913

Bando SY, Bertonha FB, Pimentel-Silva LR et al (2021) Hippocampal CA3 transcriptional modules associated with granule cell alterations and cognitive impairment in refractory mesial temporal lobe epilepsy patients. Sci Rep 11(1):10257

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113

Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 13:56–68

Becker M, Hesse V (2020) Minipuberty: why does it happen? Horm Res Paediatr 93:76–84

Berrih-Aknin S, Panse RL, Dragin N (2018) AIRE: a missing link to explain female susceptibility to autoimmune diseases. Ann N Y Acad Sci 1412:21–32

Bertonha FB, Bando SY, Ferreira LR et al (2020) Age-related transcriptional modules and TF-miRNA-mRNA interactions in neonatal and infant human thymus. PLoS One 15:e0227547

Billi AC, Kahlenberg JM, Gudjonsson JE (2019) Sex bias in autoimmunity. Curr Opin Rheumatol 31:53–61

Blondel VD, Guillaume JL, Lambiotte R et al (2008) Fast unfolding of communities in large networks. J Stat Mech P10008

Cao S, Carlesso G, Osipovich AB et al (2008) Subunit 1 of the prefoldin chaperone complex is required for lymphocyte development and function. J Immunol 181:476–484

Chaussabel D, Baldwin N (2014) Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol 14:271–280

Chen EY, Tan CM, Kou Y et al (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf 14:128

Cheng M, Anderson MS (2018) Thymic tolerance as a key brake on autoimmunity. Nat Immunol 19:659–664

Chinn IK, Blackburn CC, Manley NR et al (2012) Changes in primary lymphoid organs with aging. Semin Immunol 24:309–320

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E70:066111

Cowan JE, Takahama Y, Bhandoola A et al (2020) Postnatal involution and counter-involution of the thymus. Front Immunol 11:897

Dragin N, Bismuth J, Cizeron-Clairac G et al (2016) Estrogen-mediated downregulation of AIRE influences sexual dimorphism in autoimmune diseases. J Clin Invest 126:1525–1537

Dumont-Lagacé M, St-Pierre C, Perreault C (2015) Sex hormones have pervasive effects on thymic epithelial cells. Sci Rep 5:12895

Farooqui A, Tazyeen S, Ahmed MM et al (2018) Assessment of the key regulatory genes and their Interologs for Turner Syndrome employing network approach. Sci Rep 8:10091

Fu G, Rybakin V, Brzostek J et al (2014) Fine-tuning T cell receptor signaling to control T cell development. Trends Immunol 35:311–318

Gaiteri C, Ding Y, French B et al (2014) Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav 13:13–24

Geenen V (2021) The thymus and the science of self. Semin Immunopathol 43:5–14

Ghisi M, Corradin A, Basso K et al (2011) Modulation of microRNA expression in human T-cell development: targeting of NOTCH3 by miR-150. Blood 117:7053–7062

Gui J, Mustachio LM, Su DM et al (2012) Thymus size and age-related thymic involution: early programming, sexual dimorphism, progenitors and stroma. Aging Dis 3:280–290

Guo D, Ye Y, Qi J et al (2016) MicroRNA-181a-5p enhances cell proliferation in medullary thymic epithelial cells via regulating TGF-β signaling. Acta Biochim Biophys Sin Shanghai 48:840–849

Haljasorg U, Bichele R, Saare M et al (2015) A highly conserved NF-κB-responsive enhancer is critical for thymic expression of Aire in mice. Eur J Immunol 45:3246–3256

Hao Y, Hao S, Andersen-Nissen E et al (2021) Integrated analysis of multimodal single-cell data. Cell S0092-8674:00583–00583

Kernfeld EM, Genga RMJ, Neherin K et al (2018) A single-cell transcriptomic atlas of thymus organogenesis resolves cell types and developmental maturation. Immunity 48:1258–1270.e6

Klein SL, Flanagan KL (2016) Sex differences in immune responses. Nat Rev Immunol 16:626–638

Kondo K, Ohigashi I, Takahama Y (2019) Thymus machinery for T-cell selection. Int Immunol 31:119–125

Kuiri-Hänninen T, Sankilampi U, Dunkel L (2014) Activation of the hypothalamic-pituitary-gonadal axis in infancy: minipuberty. Horm Res Paediatr 82:73–80

Kuleshov MV, Jones MR, Rouillard AD et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44(W1):W90–W97

Laios K (2018) The thymus gland in ancient Greek medicine. Hormones (Athens) 17:285–286

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 9:559

Merrheim J, Villegas J, Van Wassenhove J et al (2020) Estrogen, estrogen-like molecules and autoimmune diseases. Autoimmun Rev 19:102468

Miller JFAP (2020) The function of the thymus and its impact on modern medicine. Science 369(6503):eaba2429

Moreira-Filho CA, Bando SY, Bertonha FB et al (2015) Community structure analysis of transcriptional networks reveals distinct molecular pathways for early- and late-onset temporal lobe epilepsy with childhood febrile seizures. PLoS One 10(5):e0128174

Moreira-Filho CA, Bando SY, Bertonha FB et al (2016) Modular transcriptional repertoire and MicroRNA target analyses characterize genomic dysregulation in the thymus of Down syndrome infants. Oncotarget 7:7497–74533

Moreira-Filho CA, Bando SY, Bertonha FB et al (2018) Minipuberty and sexual dimorphism in the infant human thymus. Sci Rep 8:13169

Nakaya HI, Wrammert J, Lee EK et al (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12:786–795

Narayanan T, Subramaniam S (2013) Community structure analysis of gene interaction networks in Duchenne muscular dystrophy. PLoS One 8:e67237

Newman MEJ (2010) Networks: an introduction. Oxford University Press, New York

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113

Obermoser G, Presnell S, Domico K et al (2013) Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. Immunity 38:831–844

Ohigashi I, Tanaka Y, Kondo K et al (2019) Trans-omics impact of thymoproteasome in cortical thymic epithelial cells. Cell Rep 29:2901–2916.e6

Park JE, Botting RA, Domínguez Conde C et al (2020) A cell atlas of human thymic development defines T cell repertoire formation. Science 367(6480):eaay3224

Passos GA, Speck-Hernandez CA, Assis AF et al (2018) Update on Aire and thymic negative selection. Immunology 153:10–20

Perniola R (2018) Twenty years of AIRE. Front Immunol 9:98

Pobezinsky LA, Etzensperger R, Jeurling S et al (2015) Let-7 microRNAs target the lineage-specific transcription factor PLZF to regulate terminal NKT cell differentiation and effector function. Nat Immunol 16:517–524

Rezzani R, Nardo L, Favero G et al (2014) Thymus and aging: morphological, radiological, and functional overview. Age (Dordr) 36:313–351

Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504

Singh Y, Garden OA, Lang F et al (2015) MicroRNA-15b/16 enhances the induction of regulatory T cells by regulating the expression of rictor and mTOR. J Immunol 195:5667–5677

Steinmann GG (1986) Changes in the human thymus during aging. Curr Top Pathol 75:43–88

Steinmann GG, Klaus B, Müller-Hermelink HK (1985) The involution of the ageing human thymic epithelium is independent of puberty. A morphometric study. Scand J Immunol 22:563–575

Thapa P, Farber DL (2019) The role of the thymus in the immune response. Thorac Surg Clin 29:123–131

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98:5116–21

van Dam S, Võsa U, van der Graaf A et al (2018) Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform 19:575–592

Varas A, Jiménez E, Sacedón R et al (2000) Analysis of the human neonatal thymus: evidence for a transient thymic involution. J Immunol 164:6260–6267

Zhao L, Zhang Y (2015) miR-342-3p affects hepatocellular carcinoma cell proliferation via regulating NF-κB pathway. Biochem Biophys Res Commun 457:370–377

Zhu M, Chin RK, Christiansen PA et al (2006) NF-kappaB2 is required for the establishment of central tolerance through an Aire-dependent pathway. J Clin Invest 116:2964–2971

Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21:1010–1024

Chapter 5
Transcriptomics at the Single Cell Level and Human Diseases: Opportunities and Challenges in Data Processing and Analysis

Vinicius Maracaja-Coutinho and Patricia Severino

5.1 Introduction

Most cells belonging to an organism share the same genome, but gene expression levels vary according to cell types and tissues. RNA amounts present in each cell are limited, so gene expression profiling has been traditionally performed with pooled cells. Despite breakthroughs in precision medicine during the past decade, bulk RNA profiling approaches conceal gene expression heterogeneity expected in samples and tissues. The recent development of single-cell RNA sequencing (scRNA-seq) enables researchers to dissect this heterogeneity through genome-wide expression profiling at cellular resolution (Kolodziejczyk and Lonnberg 2018). This information allows not only the re-evaluation of current hypothesis and biomarkers that differentiate disease subtypes and treatment subgroups, but also to distinguish cell types and cell states within tissues, possibly shedding light into molecular mechanisms underlying the pathogenesis in complex diseases.

In this chapter, we will discuss features of the most relevant technologies for single cell isolation and library preparation these days, as well as pipelines for data analysis and interpretation. We will also examine the application of scRNA-seq for biomarker discovery and challenges specific to single cell data analysis or that may also be experienced when analyzing bulk RNA sequencing data.

V. Maracaja-Coutinho
Advanced Center for Chronic Diseases – ACCDiS, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile

P. Severino (✉)
Instituto Israelita de Ensino e Pesquisa Albert Einstein – Hospital Israelita Albert Einstein, São Paulo, SP, Brazil
e-mail: patricia.severino@einstein.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_5

Although this chapter does not claim to be an exhaustive review, it should be a helpful guide for beginners in this field on state-of-the-art single-cell handling technologies and data analysis.

5.2 Isolation of Individual Cells

Single-cell RNA-seq involves isolating individual cells, then independently reverse transcribing and amplifying their mRNAs before generating barcoded libraries that are pooled for sequencing.

Numerous technologies are available for single-cell isolation and separation and they differ in mainly three aspects: levels of automation (manual to high throughput solutions), ability to isolate specific/individual cells, and compatibility with application requirements. At present, cell integrity and viability (after isolation) and single-cell yield (after separation) are technically challenging. Cell integrity is essential throughout the cell isolation process and should be kept prior to lysis to avoid early degradation of RNA for any type of application. When operating on living cells, cell viability is required and it is critical to keep in mind that cells respond to stress factors (e.g., mechanical forces and chemicals) with changes in biological processes. Thus, regardless of the selected technology for cell isolation, when working with live cells, they must be disaggregated into a suspension with rapid processing steps at near-physiological conditions since lengthy procedures will lead to undesirable alterations in gene expression or even to death of a fraction of the cell population.

Efficiency regarding yield of the single-cell isolation process becomes critical when performing single-cell analysis on rare cell types or when using costly reagents. The analysis of individually selected cells rather than of a complete population of cells from a sample imposes rigorous requirements on the separation technology, such as not only preventing aliquots that are empty, but also that no cell is lost during the isolation process. Additionally, throughput in terms of total number of single cells that can be isolated is an important factor when large cell populations with low abundance of target cells (e.g. circulating tumor cells) are the focus. In these cases, manual procedures are prohibitive and a high throughput technology must be used.

Finally, compatibility with existing workflows; acquisition, maintenance, and running costs of the platforms; and even the space needed in the laboratory for the instruments are also important criteria when selecting isolation and separation platforms.

Currently the most widely adopted technologies for cell isolation are fluorescence-activated cell sorting (FACS), limiting dilution on microplates, manual single cell picking or micromanipulation, laser capture microdissection, and microfluidics.

5.2.1 Flow Cytometry and Fluorescence-Activated Cell Sorting

The principle of flow cytometry is passing of cells, in single-file, in front of a laser beam that allows for cells to be detected, counted, and, eventually, sorted. In order for this to happen single or multiple cell components are labeled with synthetic markers such as fluorescent dyes that are excited by the laser and emit light at specific wavelengths.

Thousands of cells per second can be analyzed as they pass in front of the laser and, as fluorescent dies are measured, the amount and types of cells present in the sample are determined by means of information on their physical, chemical, or optical properties. In practice, relative size and granularity can be extracted as forward scatter (FSC) and side scatter (SSC), respectively, and a vast palette of functional properties can be measured by different fluorescent staining. There are various types of flow cytometers, but for single cells isolation, fluorescence activated cell sorting (FACS) systems are required (Herzenberg et al. 2002). In FACS systems, the bypassing cells are sorted after analysis. Briefly, cells are automatically suspended in a closed system of small channels, forced through a small nozzle forming a liquid jet that is then broken apart into a continuous stream of droplets. Electrically charged plates deflect droplets containing the cells of interest and guide them to collector tubes or micro-well plates. FACS systems provide different sort modes that specialize on high throughput, enrichment, or purity. Depending on the application, type of cells, and the chosen sorting mode, the actual rate of sorted cells per second can strongly differ between some hundred up to several thousand cells.

FACS can be suitable for rare cell sorting (subpopulations < 1%) and some of FACS systems, within minutes, are able to deposit single cells in micro-well plates with high purity. Nevertheless, drawbacks specific to this technology include the need for monoclonal antibodies to label proteins of interest, subpopulations with similar expression of labeled proteins are difficult to differentiate, large starting volumes are required, and FACS may have non-negligible effects on cell viability (Wu and Singh 2012). However, the popularity of FACS systems makes them accessible to a broad range of users.

5.2.2 Limiting Dilution on Microplates

Hand-pipettes or pipetting robots can be used to isolate individual cells from cell suspensions: the number of cells in a highly diluted sample can be as low as one single cell per aliquot due to the statistical distribution of the cells in the suspension. This procedure is termed limiting dilution, and even though it has been known for decades (Fuller et al. 2001), its application in single-cell transcriptomics may be the technology of choice depending on the application since it is a simple, gentle, and relatively cost-efficient process with reasonable throughput when using automated pipetting robots. However, due to the statistical nature of the process, further technologies such as microscopic imaging systems may be required downstream to prove single-cell presence in wells. Combined with upstream sorting or enrichment techniques it can constitute an appropriate tool to easily separate viable single cells.

5.2.3 Manual Single Cell Picking or Micromanipulation

Targeted isolation of individual single cells may be achieved by micromanipulation using a microscope -assisted picking tool. These micromanipulators for manual cell picking consist of an inverted microscope combined with movable micropipettes made of ultrathin glass capillaries coupled with an aspiration and dispensation device. For cell picking, an operator selects target cells from a suspension in a dish or well plate and performs the aspiration, transfer, and dispensation to a well of a well plate. Micromanipulators allow controlled separation of living cells and this targeted isolation of a specific cell under microscope is the central benefit of this technology, not shared by many others (Wright et al. 1998; Citri et al. 2011).

In terms of cell types it is a very flexible technology, unlike laser capture microdissection (see next section in this chapter) that mainly isolates single cells from sections of fixed tissue, micromanipulation allows the isolation of live cells, although the manual process of obtaining single cells limits the overall throughput. Furthermore, to confirm if a single cell has been successfully transferred to a well, additional observation of the well plate is required. Recent methodologies have proposed fully automated, video assisted, isolation and placement of single cells in well plates (Lu et al. 2010).

5.2.4 Laser Capture Microdissection

Individual cells or cell compartments can be isolated from solid tissue samples using laser capture microdissection (LCM) (Espina et al. 2007; Nakamura et al. 2007). Briefly, the target cell or compartment in a tissue section is visually identified through a microscope and the section to be cut off is marked with a line around it so that the laser will cut the tissue or the isolated cell or compartment along the line trajectory. Following this cutting procedure different methods can be used to extract the dissected tissue: contact-based extraction via adhesion, employing solutions such as adhesive tube caps; contact-free gravity-assisted microdissection (GAM) using an inversely mounted substrate placed over a collector tube so that once cut out by the laser, the target section falls down into the collector cube; and contact-free laser pressure catapulting (LPC), where a short defocused laser pulse ignites a local plasma below the previously cut section catapulting the section vertically against gravity into a nearby collector tube.

Whenever single cells need to be isolated from solid samples (e.g., tissue and biopsies) LCM systems are commonly the tool of choice. These systems are relatively easy to handle and the lasers cut with sub-micrometer precision without introducing deleterious heat to the tissue. However, despite a higher level of user-friendliness and automation in modern LCM systems, the selection and isolation processes remain operator based, limiting the throughput. Additionally, since the integrity of extracted cells is important for reliable downstream analysis of RNA, depending on the quality of fixation and cell extraction method used (i.e. adhesion, gravity, and catapulting) single-cell integrity might be compromised, it might remain unclear if a cell was actually transferred, and if contaminants such as fragments of adjacent cells were transferred along with the cell of interest (Fend 2000; Bevilacqua et al. 2010; Liu 2010).

5.2.5 Microfluidics

Microfluidics have been establishing new workflows for single-cell separation, isolation, and analysis (Whitesides 2006). These systems can be operated with very low volumes regarding cell samples and reagents, which is advantageous for rare cell applications or from an economical point of view, and allow the isolation of hundreds to thousands of cells with more or less automation (Svensson et al. 2018).

A widely used commercial platform is Fluidigm C1, launched in 2013. It enables automated single-cell lysis, RNA extraction, cDNA synthesis, and amplification based on Smart-seq (Ramskold et al. 2012) (see section 3.3 for sequencing details) of up to 800 single cells in parallel. Cap analysis gene expression (CAGE) can also be conducted using this system, which enables the profiling of the 5′ end of transcripts with strand information in a single cell (Kouno et al. 2019).

Droplet-based microfluidics, consisting mainly of microchannels introducing or collecting reagents and samples, allows the monodispersion of aqueous droplets in a continuous oil phase that rapidly encapsulates, in nanoliter-sized volumes, single-cell reactions (Agresti et al. 2010; Duncombe et al. 2015). The lower volume required by this system enables the manipulation and screening of thousands to millions of cells at a reduced cost. Relevant microdroplet-based systems are inDrop (1CellBio), Chromium (10× Genomics), ddSEQ (Bio-Rad/Illumina), and Nadia (Dolomite).

Finally, relevant microwell-based systems are Rhapsody (BD) and ICELL8 (Takara). These systems capture and barcode hundreds to thousands of single cells using single-cell partitioning technologies, and use visualization systems following the single-cell capture workflow to provide more control over the selection of the isolated cells.

The complexity of biological systems and the number of cells that can be processed in parallel for scRNA-seq constitute a challenge and impact statistical power. However, it is important to select appropriate methods of single cell processing according to sample type and research purposes. While droplet-based platforms offer a high throughput, meaning a lower cost per cell, platforms such as Fluidigm C1 provide lower throughput, but give the possibility to inspect the presence of the cells before lysis, ensuring the presence of a single cell for downstream analysis, saving costs in case cell isolation was inefficient. Additionally, although the Fluidigm system can analyze relatively few cells per run, as determined by its size, the platform can be used to obtain full-length cDNA libraries for each cell separately and can perform additional sequencing of libraries in user-selected wells, enabling in-depth, more sophisticated information for each cell. Chromium, on the other hand, enables the analysis of thousands of cells but the libraries of selected cells cannot be reanalyzed because they are mixed after barcoding. Thus, a choice between a small number of cells with a large amount of information for each cell (Fluidigm) and many cells with less information per cell (Chromium) should be carefully evaluated considering the study design.

5.3 Single-Cell Transcriptomics

Transcriptome analysis profiles the complete set of RNA molecules in a given biological sample. Three technologies that dominate this field have recently been extended to single cell applications: quantitative reverse transcription polymerase chain reaction (qRT-PCR), microarrays, and RNA sequencing (RNA-seq) (Esumi et al. 2008; White et al. 2011; Tang et al. 2009).

5.3.1 *Quantitative Reverse Transcription Polymerase Chain Reaction*

Based on the hybridization with fluorescent markers, qRT-PCR measures PCR product accumulation and until this date it is the most sensitive and reproducible quantification method for gene expression (Kolodziejczyk et al. 2015). However, even though recent studies have dramatically increased the number of cells profiled in a single experiment by qRT-PCR (Moignard et al. 2015), the main limitation of this technology is still its low throughput: while the number of cells measured can be increased using parallelized microfluidic approaches, the number of analyzed genes per experiment is hard to scale up since multiplexing is limited to four fluorescent dyes and, for each gene, specific primers require extensive testing (White et al. 2011; Zhong et al. 2011; Citri et al. 2011). Fundamentally, these characteristics imply high costs per cell and laborious efforts. Additionally, since qRT-PCR requires selection of target genes based on prior knowledge, it is a hypothesis-driven approach, potentially leading to a biased analysis.

5.3.2 *Microarrays*

Another popular quantification method in transcriptomics is microarrays. This technology uses pre-designed RNA probes for transcriptome-wide analyses. Even though microarrays have been used in single-cell analysis (Esumi et al. 2008; Rajan et al. 2011), they have not become a method of choice due to several limitations that include costs, limited dynamic range and sensitivity, and the requirement of relatively large amounts of RNA as starting material, a problematic feature when working with single cells.

5.3.3 Single-Cell RNA Sequencing

RNA sequencing (RNA-seq) is the most recent transcriptome measurement approach. Compared with qRT-PCR and microarrays, a major advantage of RNA-seq is the fact that it enables the unbiased profiling of the entire transcriptome. Its application to single cells was driven by new approaches for single cells isolation that have substantially increased the number of cells that can be profiled in a single experiment, as described in the previous section, including applications of Fluidigm C1 (Xin et al. 2016), and droplet-based approaches such as Drop-seq (Macosko et al. 2015), inDrop (Klein et al. 2015), and Chromium (Zheng et al. 2017).

In a typical mammalian cell only less than 5% of the total RNA is polyadenylated mRNA. Thus, for the measurements in single cells, reverse transcription (RT) and cDNA amplification will be performed from very small amounts of RNA. This means that the mRNA capture efficiency (i.e., the fraction of mRNA molecules that are actually recovered and quantified) is as important as the quantification accuracy. This is to say that contrary to bulk RNA analysis, where losses during sample preparation can be tolerated as long as the remaining sample is still representative of the original, if a large portion of a single-cell RNA sample is lost, then information on genes expressed in low copy numbers per cell will be irreversibly lost.

Diverse technologies for whole-transcriptome amplification (WTA) exist. Smart-seq is a WTA method developed for full-length cDNA amplification with oligo-dT priming and template switching (Ramsköld et al. 2012). Currently, Smart-seq2 (Picelli et al. 2013), Quartz-Seq (Sasagawa et al. 2013), and CEL-seq (Hashimshony et al. 2012) stably measure mRNAs from a single cell, with complete coverage across the genome allowing the detection of alternative transcript isoforms and SNPs, while RamDa-seq also detects non-poly(A) transcripts, long noncoding RNAs, and enhancer RNAs in single cells (Hayashi et al. 2018).

For the processing of hundreds to thousands of single cells for scRNA-seq, a number of solutions for library construction have been proposed (see section 2 of this chapter). Microdroplet- and microwell-based protocols allow easy handling of thousands of single cells and are currently popular platforms. In microdroplet-based technologies, RT is conducted with molecular/cell barcoding within oil droplets containing a cell or nucleus, a reaction liquid, and a barcoded bead. In the microwell-seq approaches a cell and a barcoded bead are isolated in a well (Han et al. 2018). Nx1-seq (Hashimoto 2019) and Seq-Well (Gierahn et al. 2017) have been reported to be portable, low-cost microwell-based platforms. Additionally, higher-throughput and lower-cost for scRNA-seq analysis are achieved with sci-RNA-seq, a combinatorial indexing method (Cao et al. 2017, 2019).

5.3.4 Biological Interpretation of scRNA-seq

Beyond computational analysis, challenges also arise when it comes to biological interpretation: how to relate molecular measurements and cellular function in health and disease. Cellular heterogeneity of human tissues has been the focus of recent scRNA-seq studies and of the international consortium Human Cell Atlas (www.humancellatlas.org) (Regev et al. 2017; Lindeboom et al. 2021; Rozenblatt-Rosen et al. 2021) that include the search for rare or new cell types in large data sets as well as contributions of different genes to each cell state. The in silico dissection of a mixture of cells into different molecular/transcriptomic states is visually interpretable by cell clusters or cell/gene networks, using tools for dimension reduction (e.g., principal component analysis or Gaussian process latent variable model), differential analysis, clustering (e.g. t-stochastic neighborhood embedding plots), or network inference (Stegle et al. 2015). Noteworthy is that as the number of cells in a given study increases due to technological improvement, so will the number of presumed transcriptomic states and, consequently, of distinct clusters detected. This scenario potentially influences the interpretation of the results and demands appropriate model selection procedures. Advances in statistical modeling of scRNA-seq data should be closely linked to proper model selection to guide heterogeneity analysis.

Computational approaches can generate hypotheses and complement but not substitute traditional experimental validation. Only from transcriptional profiles obtained from scRNA-seq data it is difficult to make conclusive statements on the functional state of cell subtypes. All exclusively big data-driven single cell analyses are mostly descriptive and lack mechanistic insights. Besides a computing infrastructure and robust statistical methods, for extracting relevant information that would guide follow-up experiments, biological expertise and a clear research question are needed.

5.4 scRNA-seq Bioinformatics and Data Analysis

The bioinformatics data analysis of scRNA-seq is much more complex than the traditional bulk RNA-seq. It requires careful execution of distinct computational steps, using stand-alone tools and R packages, together with some particular web services and specific databases. In this section, we will cover the general tools used to perform each one of the main steps in scRNA-seq data analysis, from the initial quality control and filtering of unwanted cells and transcripts, to the cell identity classification, cellular state dynamics, gene networks inference, and intercellular communication.

5.4.1 Barcoding Inspection, Reads Mapping, Quantification, and Batch Correction

First, standard FASTQ format files with sequencing reads retrieved from the sequencer must be demultiplexed into sets of cell-specific reads based on specific barcodes. Different scRNA-seq library construction platforms vary in approaches used to barcode cells and transcripts (see previous sections in this chapter), resulting in a myriad of platform-specific protocols. Some scRNA-seq approaches additionally use unique molecular identifiers (UMIs), which are tag sequences used to reduce amplification noise of transcripts within the same cell (Islam et al. 2014; Stegle et al. 2015). For instance, Chromium (10× Genomics) platform provides its own software (Cell Ranger); CELSeq2 (Hashimshony et al. 2016) barcoding was developed for Cel-seq2 protocol, but it can be used in data generated by other platforms such as SMART-seq2 and Drop-seq; and other generic tools were also made available for this purpose, such as STARsolo (Kaminow et al. n.d.), Alevin (Srivastava et al. 2019), and scPipe (Tian et al. 2018). These tools also perform the mapping of reads to the studied genome and gene expression quantification, generating cell-specific read counts expression matrices as output.

Next, the expression matrix must be evaluated by a quality control procedure, filtering out unwanted genes and cells based on several criteria. The huge amount of gene expression data from thousands of cells presents a high level of variance and noise due to cell-to-cell variation, which are caused by the particularities and limitations of the cell isolation and selection procedures, and transcripts amplification. In this process, different tools use distinct metrics to remove low-quality cells and transcripts, filtering out, for example, genes expressed in a small number of cells, as well as cells expressing a small number of genes. Seurat (Hao et al. 2021), Scater (McCarthy et al. 2017), and Scanpy (Wolf et al. 2018) workflows are widely used tools for these quality control, filtering, and bias correction steps. In this process, the proportion of reads mapping to the mitochondrial genome (mtDNA) is also evaluated as an additional quality control metric, and cells presenting an abnormal number of mapped reads to mtDNA are often removed.

5.4.2 Normalization, Feature Selection, Dimensionality Reduction, and Cell-Specific Marker Genes Identification

Once we have a high quality expression matrix consisting of read counts, the transcripts expression must be normalized to adjust for differences in experimental conditions making the expression values between cells more comparable. Examples of tools for this purpose are Scanpy (Wolf et al. 2018), Seurat (Hao et al. 2021), Monocle3 (Trapnell et al. 2014), and CORAZON (Ramos et al. 2020). It is worth mentioning that due to the particularities and 3′ biases in the library preparation of some scRNA-seq platforms, the normalization approaches that consider the transcript length in their formula (e.g., FPKM and RPKM) may not be appropriate.

In scRNA-seq experiments, one of our main goals is to characterize heterogeneity in a particular tissue and condition of interest (e.g., cell states and cell types). The clustering and dimensionality reduction are commonly used approaches to compare the expression of transcripts within cells, classifying them into groups by their expression levels. In this process, the feature selection step is essential for the selection of transcripts relevant to the biological tissue, excluding transcripts with expression patterns that do not present meaningful biological variation across the cells. Next, the selected genes must be clustered through unsupervised learning approaches to separate different groups of cells according to their expression patterns. In the clustering process, each individual transcript represents a dimension of the data, and each cell expression profile determines the location of a particular transcript in a high-dimensional space. To facilitate data processing and visualization, the number of separated dimensions of data is reduced. Finally, once the cells are separated into clusters according to their expression patterns, the cell identity is obtained through the identification of specific gene markers. All these steps can be performed by tools such as SC3 (Kiselev et al. 2017), Scanpy (Wolf et al. 2018), Seurat (Hao et al. 2021), and Monocle3 (Trapnell et al. 2014). Additionally, databases providing comprehensive information related to cell markers for different cell types and human diseases have been released, such as CellMarker (Zhang et al. 2019), PanglaoDB (Franzén et al. 2019), and CancerSEA (Yuan et al. 2019), which can be integrated into the tools to map and annotate specific cell types in the different clusters.

5.4.3 Gene Trajectories and Pseudotime

Cells in particular systems and biological conditions exhibit continuous and dynamic states, and transitions between them. Examples are the cell differentiation state of specialized cell subtypes or immune cells activation, which occur through gradual changes in their RNA expression profiles. The cell states along these processes can be computationally reconstructed through cell trajectories and pseudotime approaches, revealing key factors that could be triggering these state transitions. In summary, the cell trajectory characterizes a path through which cell populations may progress along different cellular states in a continuous process, representing the transition from an initial to a final state. In this sense, the pseudotime represents the specific state of cells along this trajectory and can reveal if a particular cell is more differentiated than the other. Widely used tools to characterize cellular state dynamics are Scanpy (Wolf et al. 2018), Monocle3 (Trapnell et al. 2014), and CellRouter (Lummertz da Rocha et al. 2018).

5.4.4 *Gene Regulatory Networks and Cell–Cell Communication Profiling*

The inference of biological networks to sets of genes expressed in specific cells and conditions of interest can provide meaningful biological insights that may not be revealed by bulk RNA-seq. However, technical noise of scRNA-seq data, the cell gene expression heterogeneity, subpopulations, and cell states may impose a level of complexity in this type of analysis. SCENIC (Van de Sande et al. 2020) and PIDC (Chan et al. 2017) were developed to reconstruct gene regulatory networks by predicting transcription factors and target transcripts associations in scRNA-seq data. CEMiTool (Russo et al. 2018) and webCEMiTool (Cardozo et al. 2019), an R package and web-server respectively, allow users to easily identify biologically relevant gene co-expression modules in an automated and easy-to-use way, as well as to perform a comprehensive set of analyses to better understand the biological functions present in the underlying system. In single-cell assays, the modules found can be used to redefine cell populations, revealing novel gene associations, and predicting gene function by a guilt-by-association approach.

Another interesting type of computational functional analysis that is gaining attention in the scRNA-seq field is the determination of cell–cell interactions and communication patterns based on single-cell data. Computational tools and databases emerged recently to decipher the intercellular signaling pathways, especially ligand–receptor pairs, based on protein–protein interaction information retrieved from scRNA-seq (Armingol et al. 2021). Examples of tools for this kind of analysis are iTALK (Wang et al. n.d.), CellChat (Jin et al. 2021), CCCExplorer (Choi et al. 2015), and ICELLNET (Noël et al. 2021).

5.5 Final Remarks

Single-cell analysis distinguishes differences between individual cells within seemingly homogeneous populations. Single-cell analysis techniques can also be applied to the detection of rare cells within heterogeneous cell populations, which would be useful in both basic research and clinical applications. Over the last decade, advanced analysis techniques have enabled the study of complex biological systems and phenomena at this single-cell resolution. However, despite the advantages of these techniques, they are not exempt from limitations and can be technically and financially demanding. This chapter addressed transcriptomics at single-cell resolution. Currently, the sensitivity of scRNA-seq is critically dependent on a variety of technical aspects that include single-cell isolation and handling, library preparation, and sequencing depth. With the number of cells tested, experimental costs increase significantly and the large amounts of complex data generated demand high levels of expertise and complex computational tools. Moreover, the implementation of scRNA-seq may require time-consuming and labor-intensive optimization of multiple sample-specific steps in their procedures. This scenario explains why single-cell analysis technologies are still not routinely performed or widespread. To enable extensive use in both laboratory and clinical settings, assays of relatively lower costs, easy to perform, and readily adaptable for a wide range of applications are awaited.

Acknowledgments VMC was funded by the following grants from Agencia Nacional de Investigación y Desarrollo (ANID), Chile: FONDECYT (1211731), and FONDAP (15130011).

References

Agresti J, Antipov E, Abate AR, Ahn K, Rowat AC, Baret JC, Marquez M, Klibanov AM, Griffiths AD, Weitz DA (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. Proc Natl Acad Sci U S A 107:4004–4009

Armingol E, Officer A, Harismendy O, Lewis NE (2021) Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet 22(2):71–88

Bevilacqua C, Makhzami S, Helbling JC, Defrenaix P, Martin P (2010) Maintaining RNA integrity in a homogeneous population of mammary epithelial cells isolated by laser capture microdissection. BMC Cell Biol 11:95

Cao J, Packer JS, Ramani V et al (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 357:661–667

Cao J, Spielmann M, Qiu X et al (2019) The single-cell transcriptional landscape of mammalian organogenesis. Nature 566:496–502

Cardozo LE, Russo PST, Gomes-Correia B, Araujo-Pereira M, Sepúlveda-Hermosilla G, Maracaja-Coutinho V, Nakaya HI (2019) webCEMiTool: co-expression modular analysis made easy. Front Genet 10(March):146

Chan TE, Stumpf MPH, Babtie AC (2017) Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst 5(3):251–267.e3

Choi H, Sheng J, Gao D, Li F, Durrans A, Ryu S, Lee SB et al (2015) Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. Cell Rep 10(7):1187–1201

Citri A, Pang ZP, Südhof TC, Wernig M, Malenka RC (2011) Comprehensive qPCR profiling of gene expression in single neuronal cells. Nat Protoc 7:118–127

da Rocha LE, Rowe RG, Lundin V, Malleshaiah M, Jha DK, Rambo CR, Li H, North TE, Collins JJ, Daley GQ (2018) Reconstruction of complex single-cell trajectories using CellRouter. Nat Commun 9(1):892

Duncombe TA, Tentori AM, Herr AE (2015) Microfluidics: reframing biological enquiry. Nat Rev Mol Cell Biol 16:554–567

Espina V, Heiby M, Pierobon M, Liotta (2007) LA Laser capture microdissection technology. Expert Rev Mol Diagn 7:647–657

Esumi S, Wu X, Yanagawa Y, Obata K, Sugimoto Y, Tamamaki N (2008) Method for single-cell microarray analysis and application to gene-expression profiling of GABAergic neuron progenitors. Neurosci Res 60:439–451

Fend F (2000) Laser capture microdissection in pathology. J Clin Pathol 53:666–672

Franzén O, Gan L-M, Björkegren JLM (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019(January). https://doi.org/10.1093/database/baz046

Fuller SA, Takahashi M, Hurrell JG (2001) Cloning of hybridoma cell lines by limiting dilution. Curr Protoc Mol Biol:Chapter 11:Unit11.8

Gierahn TM, Wadsworth M, Hughes T et al (2017) Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods 14:395–398

Han X, Wang R, Zhou Y et al (2018) Mapping the mouse cell atlas by microwell-seq. Cell 172:1091–1107.e17

Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ et al (2021) Integrated analysis of multimodal single-cell data. Cell. https://doi.org/10.1016/j.cell.2021.04.048

Hashimoto S (2019) Nx1-Seq well based single-cell analysis system. Adv Exp Med Biol 1129:51–61

Hashimshony T, Wagner F, Sher N, Yanai I (2012) CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. Cell Rep 2:666–673

Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, Gennert D et al (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol 17(April):77

Hayashi T et al (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun 9:619

Herzenberg LA, Parks D, Sahaf B, Perez O, Roederer M, Herzenberg LA (2002) The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clin Chem 48:1819–1827

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S (2014) Quantitative single-cell RNA-Seq with unique molecular identifiers. Nat Methods 11(2):163–166

Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan C-H, Myung P, Plikus MV, Nie Q (2021) Inference and analysis of cell-cell communication using CellChat. Nat Commun 12(1):1088

Kaminow B, Yunusov D, Dobin A (n.d.) STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-Seq data. https://doi.org/10.1101/2021.05.05.442755

Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN et al (2017) SC3: consensus clustering of single-cell RNA-Seq data. Nat Methods 14(5):483–486

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201

Kolodziejczyk AA, Lönnberg T (2018) Global and targeted approaches to single-cell transcriptome characterization. Brief Funct Genomics 17:209–219

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. Mol Cell 58:610–620

Kouno T, Moody J, Kwon ATJ et al (2019) C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nat Commun 10:360

Lindeboom RGH, Regev A, Teichmann SA (2021) Towards a human cell atlas: taking notes from the past. Trends Genet 37:625–630

Liu A (2010) Laser capture microdissection in the tissue biorepository. J Biomol Tech 21:120–125

Lu Z, Moraes C, Ye G, Simmons CA, Sun Y (2010) Single cell deposition and patterning with a robotic system. PLoS One 5(10):e13542

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214

McCarthy DJ, Campbell KR, Lun ATL, Wills QF (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-Seq data in R. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw777

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa SI, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens

B (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol 33:269–276

Nakamura N, Ruebel K, Jin L, Qian X, Zhang H, Lloyd RV (2007) Laser capture microdissection for analysis of single cells. Methods Mol Med 132:11–18

Noël F, Massenet-Regad L, Carmi-Levy I, Cappuccio A, Grandclaudon M, Trichot C, Kieffer Y, Mechta-Grigoriou F, Soumelis V (2021) Dissection of intercellular communication using the transcriptome-based framework ICELLNET. Nat Commun 12(1):1089

Picelli S et al (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods 10:1096–1100

Rajan S, Djambazian H, Dang HCP et al (2011) The living microarray: a high-throughput platform for measuring transcription dynamics in single cells. BMC Genomics 12:115

Ramos TAR, Maracaja-Coutinho V, Miguel Ortega J, do Rêgo TG (2020) CORAZON: a web server for data normalization and unsupervised clustering based on expression profiles. BMC Res Notes 13(1):338

Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, Sandberg R (2012) Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30:777–782

Regev A, Teichmann SA, Lander ES et al (2017) The human cell atlas. elife 6:e27041

Rozenblatt-Rosen O, Shin JW, Rood JE, Hupalowska A, Regev A, Heyn H (2021) Building a high-quality human cell atlas. Nat Biotechnol 39(2):149

Russo PST, Ferreira GR, Cardozo LE, Bürger MC, Arias-Carrasco R, Maruyama SR, Thiago D. C. Hirata, et al. (2018) CEMiTool: a bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinf 19(1):56

Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR (2013) Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. Genome Biol 14:R31

Srivastava A, Malik L, Smith T, Sudbery I, Patro R (2019) Alevin efficiently estimates accurate gene abundances from dscRNA-Seq data. Genome Biol 20(1):65

Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 16:133–145

Svensson V, Vento-Tormo R, Teichmann SA (2018) Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 13:599–604

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382

Tian L, Su S, Dong X, Amann-Zalcenstein D, Biben C, Seidi A, Hilton DJ, Naik SH, Ritchie ME (2018) scPipe: a flexible R/bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLoS Comput Biol 14(8):e1006361

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014) The Dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32(4):381–386

Van de Sande B, Flerin C, Davie K, De Waegeneer M, Hulselmans G, Aibar S, Seurinck R et al (2020) A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc 15(7):2247–2276

Wang Y, Wang R, Zhang S, Song S, Jiang C, Han G, Wang M, Ajani J, Futreal A, Wang L (n.d.) iTALK: an R package to characterize and illustrate intercellular communication. https://doi.org/10.1101/507871

White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski MMA, Piret J, Aparicio S, Hansen CL (2011) High-throughput microfluidic single-cell RT-qPCR. Proc Natl Acad Sci U S A 108:13999–14004

Whitesides GM (2006) The origins and the future of microfluidics. Nature 442:368–373

Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19(1):15

Wright G, Tucker MJ, Morton PC, Sweitzer-Yoder CL, Smith SE (1998) Micromanipulation in assisted reproduction: a review of current technology. Curr Opin Obstet Gynecol 10:221–226

Wu M, Singh AK (2012) Single-cell protein analysis. Curr Opin Biotechnol 23:83–88

Xin Y, Kim J, Ni M, Wei Y, Okamoto H, Lee J, Adler C, Cavino K, Murphy AJ, Yancopoulos GD, Lin HC, Gromada J (2016) Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. Proc Natl Acad Sci U S A 113:3293–3298

Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, Xu L et al (2019) CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res 47(D1):D900–D908

Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T et al (2019) CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 47(D1):D721–D728

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049

Zhong Q, Bhattacharya S, Kotsopoulos S, Olson J, Taly V, Griffiths AD, Link DR, Larson JW (2011) Multiplex digital PCR: breaking the one target per color barrier of quantitative PCR. Lab Chip 11:2167–2174

Chapter 6
Methods for Gene Co-expression Network Visualization and Analysis

Carlos Alberto Moreira-Filho, Silvia Yumi Bando, Fernanda Bernardi Bertonha, Filipi Nascimento Silva, and Luciano da Fontoura Costa

6.1 Introduction

The development of high-throughput techniques for concurrently measuring the expression levels of thousands of genes, mostly based on DNA microarrays (Bumgarner 2013) or on RNA sequencing (RNA-seq) techniques (Cockrum et al. 2020), allowed monitoring cell's transcriptional activity across multiple conditions, opening broad perspectives for functional genomics (Joshi et al. 2021). Hereafter, new insights were gained on the genomic mechanisms underlying relevant biological processes – such as cell cycle and development, and the genome–environment interplay – leading to a systemic approach for the identification of disease-related genes and their interaction with external stressors (Vermeulen et al. 2021). A large part of this remarkable progress required multi-omics data integration (genomics, transcriptomics, proteomics, metabolomics, etc.) using systems biology computational tools (Manzoni et al. 2018; Vlachavas et al. 2021). Essentially, systems biology aims to explain biology in terms of interacting components (Joshi et al. 2021). Gene co-expression network analysis is a systems biology method for describing the correlation patterns among genes across DNA microarray or RNA-seq samples that heavily relies on network science (Gysi and Nowick 2020), as shown in the subsequent sections of this chapter.

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-87821-4_6) contains supplementary material, which is available to authorized users.

C. A. Moreira-Filho (✉) · S. Y. Bando · F. B. Bertonha · F. N. Silva · L. da Fontoura Costa
Departamento de Pediatria, Faculdade de Medicina da Universidade de São Paulo,
São Paulo, SP, Brasil
e-mail: carlos.moreira@hc.fm.usp.br; silvia.bando@hc.fm.usp.br;
fernanda.bernardi@fm.usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_6

Although RNA-seq is becoming increasingly popular for measuring gene expression, DNA microarray technology remains widely used for transcriptomics, since it is cheaper and easier to analyze (Mantione et al. 2014; Costa-Silva et al. 2017; Rao et al. 2019; Geraci et al. 2020). A single DNA microarray generates data on the expression levels of thousands of genes and typical microarray studies encompass multiple arrays covering several distinct experimental conditions, e.g., tissue samples from patients and controls, or cultured cells submitted to different treatments (Moreira-Filho et al. 2016; Bando et al. 2019). Advanced statistical and computational tools have been developed to deal with the large amount of data derived from microarray experiments (Zhang and Horvath 2005; Lee and Tzou 2009; Faro et al. 2012; Ang et al. 2016; Arunkumar et al. 2017), including machine learning techniques (Mahendran et al. 2020). One of the most effective methods for the analysis of microarray data is based on the construction of gene co-expression networks, or GCNs: gene expression levels are pairwise compared and the pairs above a cutoff threshold are linked to create a gene–gene interaction network (Weirauch 2011). The topological and dynamic properties of these networks provide important clues for understanding the functional organization of cells and tissues (Barabási and Oltvai 2004; Zhu et al. 2007; Gaiteri et al. 2014, van Dam et al. 2018). This chapter is centered in network-based methods for analyzing DNA microarray data: the fundamentals for construction, visualization, interpretation, and validation of GCNs will be discussed in the next paragraphs, emphasizing the use of graph methods (Barabási and Oltvai 2004; Barabási et al. 2011; Costa et al. 2011; Villa-Vialaneix et al. 2013; Winterbach et al. 2013). Because the identification of transcriptional modules (sets of highly interconnected genes) is of utmost biological importance, there is a specific section on Weighted Gene Co-expression Network Analysis (WGCNA), a bioinformatics tool for identifying relationships between transcriptional modules, and between these modules and phenotypic and/or clinical traits (Langfelder and Horvath 2008; Galán-Vásquez and Perez-Rueda 2019; Bando et al. 2021).

A few considerations are still needed before we start to review the network-based approach to functional genomics. First, one should keep in mind that gene function is not isolated: the network effect of genes is the driving force moving cell metabolism from one steady state to another, frequently in response to environmental changes (Sieberts and Schadt 2007; Liu et al. 2012a; Sahni et al. 2013). These transitions shape what we call complex phenotypes – normal or altered by a disease state – and can be correlated with specific changes in GCNs (Benson and Breitling 2006; Carter et al. 2013; Bando et al. 2019). Second, for studying these network changes are mandatory: (i) to gain access to the cells or tissues specifically involved in the physiological or pathological process under investigation; (ii) to collect an adequate number of biological replicates; (iii) to obtain good quality RNA samples; and (iv) to use a microarray (or RNA-seq) platform suitable for attaining the research goals. Therefore, comments on sample quality and experimental design will be made in the following section.

6.2 Analysis of DNA Microarray Data

The DNA microarray for assaying gene expression (Bumgarner 2013) consists in grid that can contain tens of thousands of probes corresponding to known transcripts of a particular genome (human, rat, etc.). Fluorescent-labeled complementary DNA (DNA synthesized from messenger RNA) samples are hybridized to probes and the relative abundance of each sequence in a sample is quantified in microarray scanner for fluorescence detection (image capture). The steps from RNA extraction to array scanning, data export, and subsequent statistical and network analyses are outlined in the following subsections. This workflow is presented in Fig. 6.1.

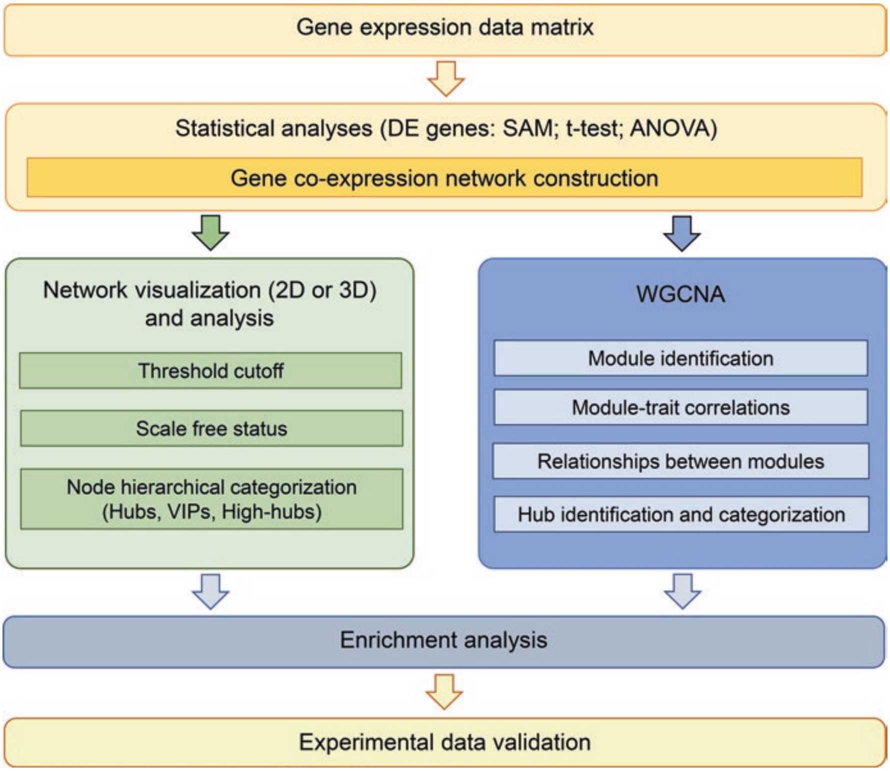

Gene expression data matrix
Statistical analyses (DE genes: SAM; t-test; ANOVA)
Gene co-expression network construction
Network visualization (2D or 3D) and analysis
Threshold cutoff
Scale free status
Node hierarchical categorization (Hubs, VIPs, High-hubs)
WGCNA
Module identification
Module-trait correlations
Relationships between modules
Hub identification and categorization
Enrichment analysis
Experimental data validation

Fig. 6.1 Workflow for gene co-expression network construction, visualization, and analysis. The diagram depicts the following steps: gene expression data matrix; statistical analysis and gene co-expression network construction; network visualization and analysis (2D or 3D); WGCNA; enrichment analysis; experimental data validation

6.2.1 RNA Isolation and Preservation

Messenger RNA for DNA microarray experiments must be preserved in its last physiological state and be prevented from degrading. Larger tissue samples (>50 mg) may be snap-frozen in liquid nitrogen. Smaller tissue samples (5-mm thick fragments) are usually preserved in RNAlater®, a product (distributed by Ambion and Qiagen) which penetrates cell membranes and inactivates RNAses. After RNA extraction, RNA quality should be assessed in a microfluidics-based platform (e.g., BioAnalyzer) for sizing, quantification, and quality control. The integrity of RNA molecules is estimated by using the RIN algorithm (Schroeder et al. 2006). RIN values range from 1 (total degradation) to 10 (intact). As a rule, only RNA samples with RIN values of 7 or higher should be used in DNA microarray experiments. Irrespective of the cellular RNA extraction protocol adopted, a final column purification step (e.g., RNeasy) consistently leads to a high yielding synthesis of cDNA.

6.2.2 Gene Expression Analysis

Scanner generated data (image file) are pre-processed, filtering out probes flagged as unreliable (low intensity, saturation, restriction control probes, etc.) by the scanning software, and thereafter normalized ending up with a file of numerical values corresponding to probe's expression levels in a microarray experiment. The assessment of raw data quality and data grouping (comparison groups, e.g., patients and controls) can be done using free software packages, like R software (R Development Core Team 2012), for normalization (Lowess test for arrays normalization), outlier exclusion, and exporting of valid transcript expression data.

MeV (TIGR Multiexperiment Viewer) is a popular free software for comparative analysis that can be used for clustering, visualization, classification, statistical analysis, and biological theme (Gene Ontology, or GO) discovery (Saeed et al. 2003). The differentially expressed transcripts for two comparison groups are obtained using SAM test – Significance Analysis for Microarray – for parametric analysis (using non-parametric statistics) or Wilcoxon–Mann–Whitney test for non-parametric analysis. ANOVA is used for multiple comparisons across conditions. Thereafter, false discovery rate tests are applied (already included in SAM test). Finally, the differential GO annotated gene expression data can be used for gene expression analyses (Fig. 6.1) and in the construction of co-expression networks, as described in Sect. 6.3.

6.3 Construction and Analysis of GCNs

GCNs can be obtained for a subset of genes, i.e., differentially expressed GO annotated genes (DE networks), or for all valid GO annotated genes (complete, or CO networks). These networks are constructed based on gene–gene covariance correlation, usually using Pearson's or Spearman's rank correlation (Fig. 6.1) (Prifti et al. 2008; Song et al. 2012). Genes presenting similar patterns of expression are strongly bounded together forming a weighted complete graph.

In order to construct a GCN links are removed from the initially complete graph by gradually increasing the correlation threshold (Elo et al. 2007). After link strength threshold adoption, usually above 0.80, the network is tested for scale-free status (see Sect. 6.3.2) by Kolmogorov–Smirnov (K-S) statistics, i.e., power law distributions in empirical data (Clauset et al. 2009). Here we used a demonstrative example of a "patient versus control" gene co-expression analysis (Fig. 6.2). This analysis considered: 202 genes and 561 links for patients' DE network; 219 genes and 486 links for control DE networks; 6,927 genes and 12,768 links for patients' CO networks; 6,705 genes and 12,468 links for control CO networks. Link strength cutoffs were 0.998 for control CO network and 0.999 for the other three networks. Figures 6.2 and 6.3 show K-S distribution for DE and CO networks, respectively, of patients' group (Figs. 6.2c and 6.3c) and controls' group (Figs. 6.2d and 6.3d).

The number of samples available for each gene is also directly connected to the statistical significance of the generated GCN. Networks constructed from datasets with less than five samples per gene can lead to high adherence to the null model, where nodes are randomly connected, thus presenting degree distributions with asymptotic exponential decay behavior. This effect can occur even when considering a very large correlation threshold, such as above 0.999.

6.3.1 Network Visualization

The Cytoscape free software (Saito et al. 2012; www.cytoscape.org) is very useful for data analysis and visualization of DE networks or subnetworks (Fig. 6.2a, b). On the other hand, CO network analysis is only possible through 3D visualization (Fig. 6.3a, b and Videos 6.1 and 6.2). Several 3D visualization softwares for gene–gene and protein–protein networks are being developed (Ishiwata et al. 2009; Pavlopoulos et al. 2008; Wang et al. 2013). One of them – developed by Luciano Costa's Research Group, Institute of Physics at São Carlos, University of São Paulo (Bando et al. 2013), and suitable for obtaining visualization of large complex networks – is based on the Fruchterman–Reingold algorithm, FR (Fruchterman and Reingold 1991), which is a force-directed technique based on molecular dynamics employing both attractive and repulsive forces between nodes (Silva et al. 2013).

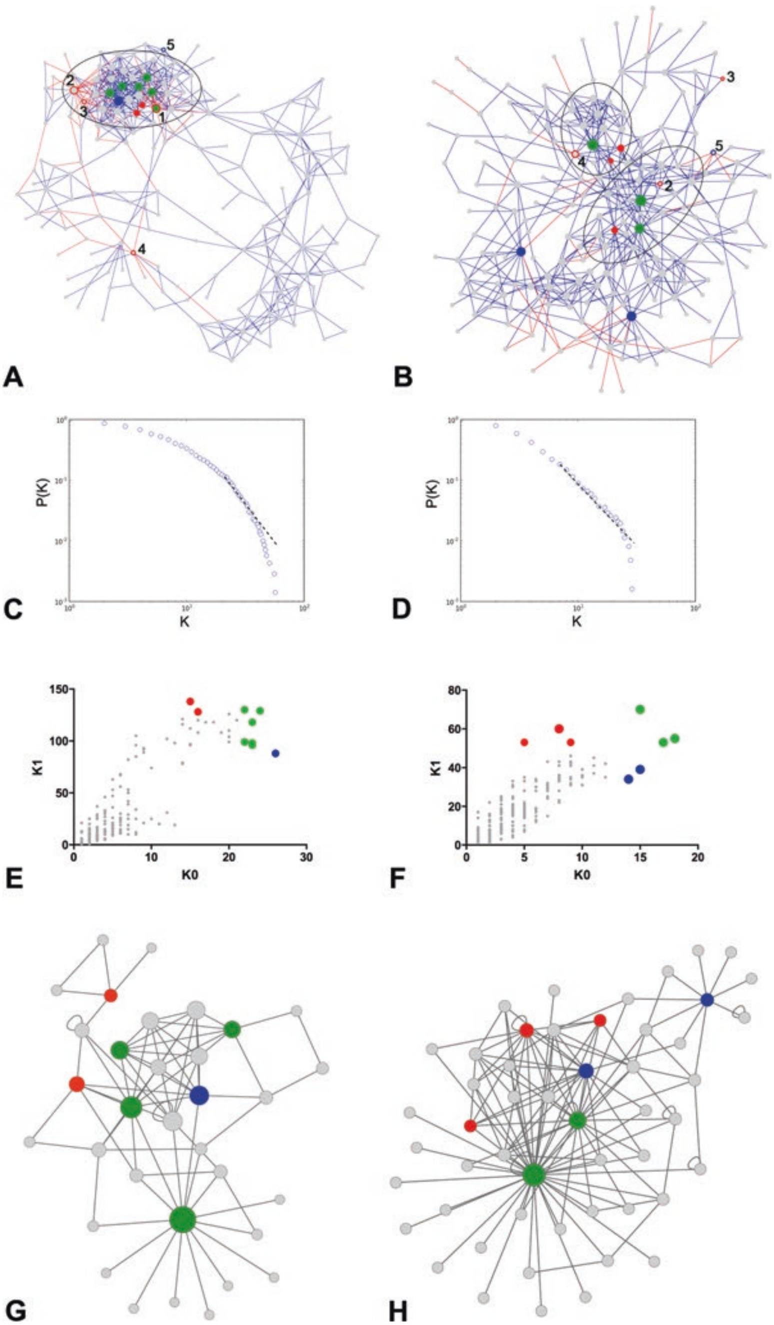


Fig. 6.2 Comparative DE network analysis for patients and control groups in a hypothetical network. DE co-expression networks for patients (**a**) and control (**b**) groups; links in blue or red indicate positive or inverse covariance correlation. It is interesting to note that the same genes (numbered nodes bordered in red or blue) have different covariance correlation between patients and control groups. Clusters are encircled in (**a** and **b)** networks. (**c** and **d**) Kolmogorov–Smirnov test for scale-free status for patients and control groups, respectively. Scatterplot of node degree (k_0) *vs* concentric node degree (k_1) measures for patients (**e**) and control (**f**) groups. Interactome in silico validations for patients and control networks are depicted in (**g** and **h)**, respectively. Hubs, VIPs, and high-hubs are indicated in blue, red, and green, respectively. Network analyses and visualization were accomplished through Cytoscape

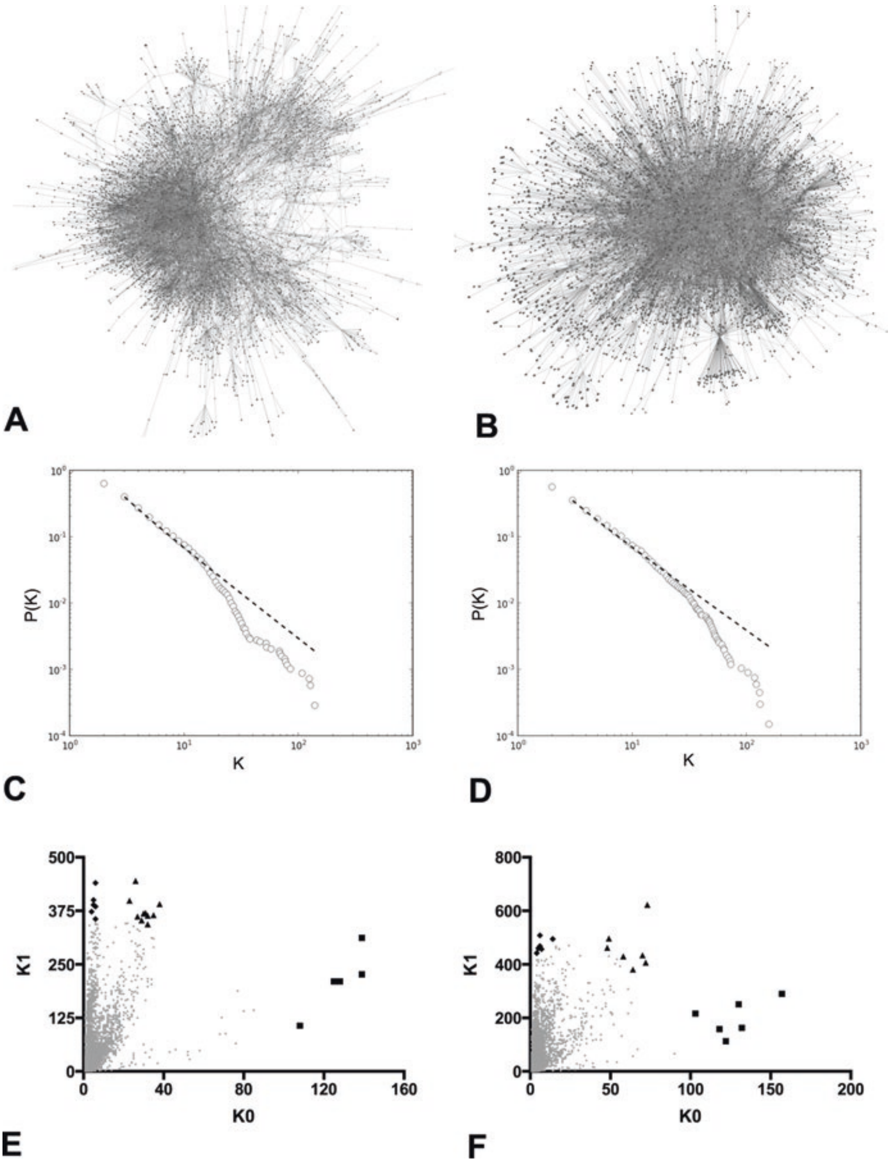


Fig. 6.3 Complete CO networks analysis for patients and control groups. CO co-expression networks for patients (**a**) and control (**b**) groups. Kolmogorov–Smirnov test for scale-free status for patients (**c**) and control (**d**) groups. Scatterplot of node degree (k0) *vs* concentric node degree (k1) measures for patients (**e**) and control (**f**) groups. Hubs, VIPs, and high-hubs are indicated by rectangles, diamonds, and triangles, respectively. For 3D CO network visualization access the video hyperlinks (Videos 6.1 and 6.2)

6.3.2 GCNs Are Scale-Free Networks

GCNs, like other biological networks and similarly to social and internet networks, are not random and follow some basic principles (Newman 2010). In random networks the nodes have nearly the same number of links and, therefore, highly linked nodes are rare. In network terminology, the number of links, or edges, connected to a node is called node degree. Hence, nodes in random networks characteristically have low diversity of node degrees. Conversely, most of the "real world networks," as GCNs or protein–protein networks, are scale free, what means that the degree distribution follows a power law: the node degree distribution P(k), with node degree k, follows P(k) ~ $k^{-\gamma}$, where γ is the degree exponent. Therefore, scale-free networks have a limited number of highly connected nodes, or hubs, that, as we shall discuss latter, are usually associated with relevant biological functions and responsible for the network robustness, i.e., hold the whole network together (Winterbach et al. 2013).

The categorization of nodes according to their node degree encompasses two other categories besides the hubs. The VIPs (a term coined in the study of social networks) are nodes presenting low node degree but connected only with hubs (Masuda and Konno 2006; Mcauley et al. 2007). In some networks VIPs may represent the highest control hierarchy in a system and hubs may be under VIPs' influence. Some nodes may present VIP status (connected with many hubs) and also present high overall number of connections, being called high-hubs (Bando et al. 2013). These hierarchical categories are all coherent with the biological role and dynamic behavior of GCNs hubs, as discussed below.

Some hubs are highly interlinked in local regions of a network thereby forming network clusters, topologically called modules or communities. Modules may be associated with specific biological processes in gene co-expression and protein–protein networks. For this reason, hubs may be sometimes classified as "party hubs", those functioning inside a module, or "date hubs", i.e., those linking different processes and organizing the network, playing a role similar to VIPs and high-hubs (Zhu et al. 2007; Barabási et al. 2011; Weirauch 2011).

6.3.3 Concentric Characterization of Nodes

One way to classify network nodes as VIPs, hubs, or high-hubs is by obtaining the node degree, k_0, and the first level concentric node degree, k_1, which takes into account all node connections leaving from its immediate neighborhood, then projecting all node values in a k_0 vs k_1 graphic. VIPs should present low k_0 but high k_1, while hubs present high k_0 and low k_1, and high-hubs present high k_0 and k_1 values. Figures 6.2e–f and 6.3e–f show each of these node categories in scatterplots of node degree vs concentric node degree measures obtained in DE and CO networks distribution scatterplots (k_0 vs k_1) generated for distinct GCNs.

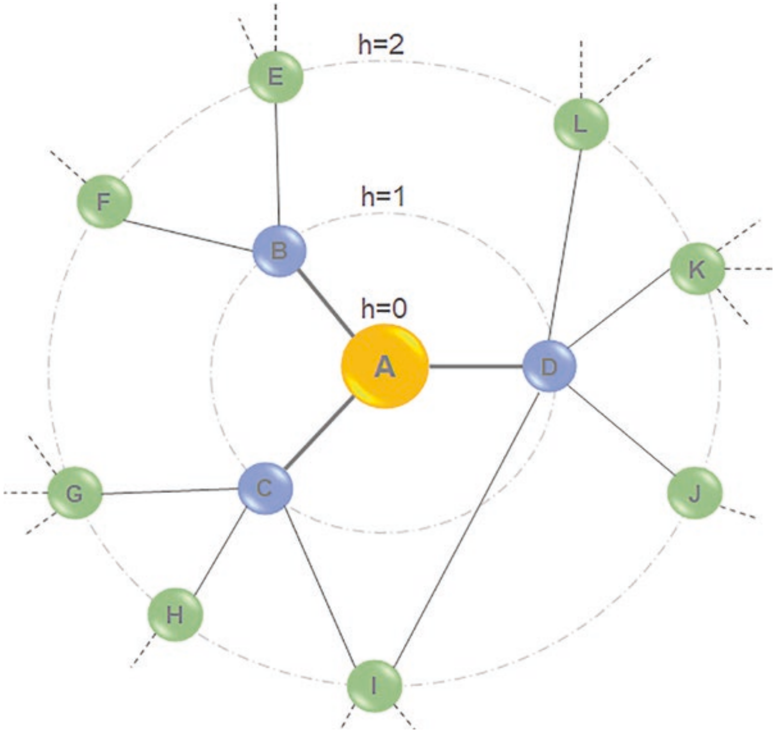


Fig. 6.4 Concentric levels. Example of concentric levels of a network for node A as reference (i.e., centered at node A). Each concentric level is represented by rings $R_h(A)$, namely $R_0(A) = \{A\}$, $R_1(A) = \{B, C, D\}$, and $R_2(A) = \{E, F, G, H, I, J, K\}$, with concentric node degrees $k_0(A) = 3$ and $k_1(A) = 8$

Because most real networks present scale-free distributions, there is no clear definition for setting a degree threshold for which we can classify nodes as being hubs or not (Barabási and Oltvai 2004; Barabási et al. 2011). This same is true for objectively defining VIPs and high-hubs, since the distribution of k_1 also suffers from the problem of not presenting a scale. Here we define hubs, VIPs, and high-hubs by ranking them according to k_0 and k_1, and then considering a set of those presenting the highest values of each property, as depicted in Fig. 6.4 (see also Bando et al. 2013). These measures are used for nodes categorization such as Hub (high k_0 VIP (high k_1 and low k_0)) and High-hub (high k_0 and k_1).

6.3.4 Betweenness Centrality

Betweenness centrality (Costa et al. 2008; Freeman 1978; Brandes 2001) is a measurement of node importance which considers the entire set of shortest paths between nodes and passing through a particular node in a network. Betweenness is one of the most important topological properties of a network: nodes with the highest betweenness control most of the information flow in the network (Yu et al. 2007; Azevedo and Moreira-Filho 2015; van Dam et al. 2018).

6.3.5 Positive or Inverse Gene–Gene Correlation

Pearson's correlation coefficient (PCC) gives us the strength of the relationship between a pair of genes (nodes in the network) (Allen et al. 2010). PCC ranges from −1 to 1 and the closer the number to either of these boundaries, the stronger the relationship: a negative number indicates an inverse correlation (e.g., expression of gene A increases as expression of gene B decreases) while a positive number indicates a positive correlation (e.g., as A increases B tends to increase). This is depicted by the blue (positive correlation) and red (negative correlation) edges in Fig. 6.2a, b.

6.3.6 Network Connectivity

This and the two next subsections will address issues on network topology. Network topology exerts a pivotal role in unraveling GCNs organization and performance under different conditions (Barabási and Oltvai 2004; Zhu et al. 2007; Costa et al. 2011; Liu et al. 2012b; Bando et al. 2019). Network connectivity is an elementary network property: a pair of nodes that have just one independent path between them are weakly connected than a pair that has many paths (Flake et al. 2002). Connectivity is commonly visualized as bottlenecks between nodes and formalized by the notion of cut set (Newman 2010). A node cut set is a set of nodes whose removal will disconnect a specific pair of nodes. Conversely, an edge cut set (or link cut set) is a set of edges whose removal will disconnect a pair of edges. A node with a higher degree of links (edges) is better connected in the network and it is supposed to play a more important role in maintaining the network structure (Barabási and Oltvai 2004; Albert 2005), what is generally associated with a relevant biological role (Langfelder et al. 2013). Connectivity is the most widely used concept for distinguishing the nodes of a network (Horvath and Dong 2008). Densely interconnected groups of nodes, or clusters (pointed out by arrows in Fig. 6.2a, b and in color in Videos 6.1 and 6.2), are frequently found in most GCNs and protein–protein networks (Newman 2006) in accordance with its scale-free connectivity distribution (Winterbach et al. 2013). These groups form topological modules, i.e., highly interlinked regions in a network, and have been associated, in GCNs and protein–protein networks, with highly conserved genes (Barabási and Oltvai 2004) and genes involved with complex diseases (Tuck et al. 2006; Cai et al. 2010; Barabási et al. 2011).

Robustness of complex networks is associated with the capacity of a network to preserve its topological features, such as connectivity and average path length, after the removal of a set of nodes or edges. Scale-free networks such as GCNs are found to be very resilient to random node/edge attacks. This means that random failures or perturbations in some nodes or sub-mechanisms do not seem to drive the entire system to a critical condition (Albert et al. 2008). However, attacks targeting nodes with high number of connections, i.e., hubs, present the opposite effect, thus removing a small number of such nodes in scale-free networks causes a huge impact on the network diameter and on its functionality performance (Azevedo and Moreira-Filho 2015).

6.3.7 Network Motifs

In biological networks it is possible to identify groups of nodes that link to each other forming a small subnetwork, or subgraph, at numbers that are significantly higher than those in randomized networks (Milo et al. 2002). These subgraphs are called motifs. Network motifs constitute smaller common patterns, or “building blocks”, of GCNs (Barabási and Oltvai 2004; Weirauch 2011) and were found to be associated with some optimized biological functions, such as feedback and feedforward loops, related to transcriptional regulation (Shen-Orr et al. 2002; Zhang et al. 2007; Watkinson et al. 2009). Molecular components of a particular motif frequently interact with nodes in outside motifs, and aggregation of motifs into motif clusters is likely to occur in many real networks (Ravasz et al. 2002). As pointed out by Barabási and Oltvai (2004), because “the number of distinct subgraphs grows exponentially with the number of nodes that are in a subgraph, the study of larger motifs is combinatorially unfeasible.” The alternative is to identify groups of highly connected nodes, called modules, directly from the network topology and manage to correlate these topological entities with their functional role (Winterbach et al. 2013).

6.3.8 Network Modules

Modules are large subgraph units, encompassing groups of densely associated nodes and connected to each other with loose links: in GCNs, for instance, modules may be hub clusters tenuously connected by VIPs. Modules serve to identify gene functions in a GCN and – as it was already observed for protein networks (Yu et al. 2007; Zhu et al. 2007) – contain “module organizer” genes, highly connected to other genes (equivalent to hubs and high-hubs) and essential to module functioning, and “connector” genes, linking different modules and relevant for intermodule communication (equivalent to VIPs) (Weirauch 2011; Bando et al. 2013, Moreira-Filho et al. 2015).

There are many statistical and computational methods for identifying modules in scale-free networks. One of them, the Girvan–Newman algorithm (Girvan and Newman 2002), is centered on defining the boundaries of modules by searching for those edges with high betweenness, i.e., more likely to link different modules. This is an important issue: cell functions are carried out in a very modular way. Modular structure reflects a group of functionally linked nodes (genes) acting together to accomplish a specific task: it may be invariant protein–RNA complexes involved post-transcriptional control of RNAs, or temporally coregulated genes controlling processes such as cell cycle and differentiation, or bacterial response to growth and stress conditions (Costanzo et al. 2010; Wang and Zheng 2012; Rosenkrantz et al. 2013, Moreira-Filho et al. 2016; Bando et al. 2017).

6.3.9 GCNs Are Modular Scale-Free Networks

The GCNs have a hub-dominated architecture, containing modules, or clusters, constituted by a highly connected number of nodes. The clustering coefficient C is a measure of the degree to which nodes in a graph tend to cluster together (Watts and Strogatz 1998). The average clustering coefficient < C > is significantly higher in most biological networks (gene–gene, protein–protein) than in random networks of equivalent size and distribution (Barabási and Oltvai 2004). In Fig. 6.2a, b GCN gene clusters appear encircled by a solid line and in Videos 6.1 and 6.2 the clusters are identified by distinct colors.

Network modules, or clusters, are present in all cellular networks and identifiable by clustering methods based on network's topology description (Newman 2006; Li and Horwath 2009) or by combining topology and functional genomics data (Wang and Zheng 2012; Weiss et al. 2012). Therefore, finding out correspondences between cluster topology and functional properties is the main goal of GCN analysis. A large amount of evidence show that modules involved in closely related biological functions tend to interact and are proximally located in the network (reviewed in Barabási et al. 2011). As we mentioned before (Sect. 6.3.6), scale-free networks are robust but attacks targeting highly connected nodes may cause network disruption. There are now compelling data linking the establishment of complex diseases with the perturbation (by mutation or altered expression) of highly connected genes in GCNs (Barabási et al. 2011; Cho et al. 2012; Liu et al. 2012a; Sahni et al. 2013; Gaiteri et al. 2014; Sahni et al. 2015; Moreira-Filho et al. 2015; Hu et al. 2016; Bando et al. 2019). Thus, functional and disease modules overlap and the transition between health and disease can be described as a module breakdown.

Another challenging issue is to understand network controllability. Controllability analysis in complex networks, a concept introduced by Liu and Barabási (Liu et al. 2011), determines the minimum set of driver nodes necessary to (linearly) control an entire system. This also allows determining the degrees of freedom that a system can attain; therefore, it can also be understood as a measurement of the network complexity. Shortly thereafter, Liu et al. (2012b) introduced the control centrality measurement which considers the individual control potential of each node in a system. As GCNs may represent complex control systems, this new framework can be helpful to understand its control hierarchical structure. However, the inference of causality, i.e., who controls whom, still presents as an open problem in network theory (Wu et al. 2012, 2019; Yuan et al. 2013).

6.4 Weighted Gene Co-expression Network Analysis (WGCNA)

The Weighted Gene Co-expression Network Analysis (WGCNA) package (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/) is a comprehensive collection of R functions (R Core Team 2012) that renders possible to identify and characterize gene modules whose members share strong co-expression. (Langfelder and Horvath 2008; Zhao et al. 2010; Pei et al. 2017; Galán-Vásquez and Perez-Rueda 2019; Kakati et al. 2019). WGCNA is based on the concept of scale-free network. It is assumed that all genes are connected. The connection strength is quantified by measures of gene expression correlation. Hence, the relative importance of a gene in a network is defined by its connectivity (Zhao et al. 2010).

Pearson's correlation coefficient is applied to calculate correlation patterns among genes across all samples or conditions and for the subsequent construction of an adjacency matrix using soft power and topological overlap matrix (TOM). Soft-thresholding process transforms the correlation matrix to mimic the scale-free topology. Module identification is based on TOM and in average linkage hierarchical clustering. Keeping to the scale-free topology criterion, a soft power β is considered. Finally, Dynamic Tree Cut algorithm is used for dendrogram's branch selection. The module eigengene (ME) is defined as the first principal component of a given module, which can be considered a representative of the gene expression profiles in a module. Module Membership (MM), also known as eigengene-based connectivity (kME), is defined as the correlation of each gene expression profile with the module eigengene of a given module (Langfelder and Horvath 2008).

WGCNA is one of the most widely used co-expression network techniques and its underlying methods can be used in both microarray gene expression data and RNA-Seq data (Kakati et al. 2019). The identification of transcriptional modules (Chaussabel and Baldwin 2014) and their association with specific phenotypic traits allow the identification of highly connected genes (Hubs) and high gene significance (HGS) genes for the traits of interest.

6.4.1 Module-Trait Association

Gene significance (GS) stands for the value of the correlation between the gene expression and particular traits. The mean GS considered for a module is the measure of the module significance (MS). The GS values are obtained using Pearson's

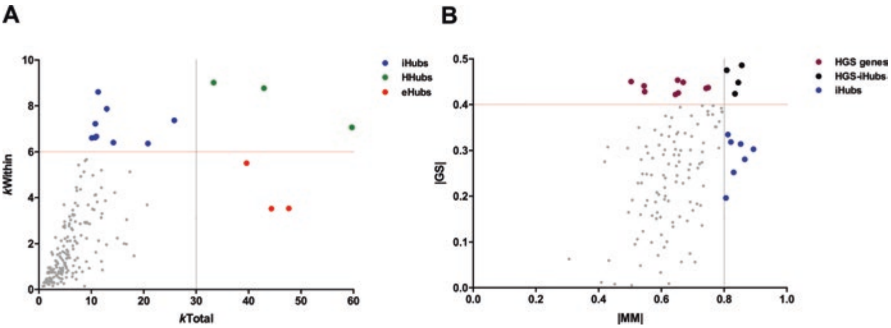


Fig. 6.5 Scatterplots for selection of modular highly connected genes (Hubs) and high gene significance (HGS) genes for the trait of interest, according to WGCNA. *k*Total *vs*. *k*Within plot for intramodular Hub selection (**a**), where iHubs, eHUbs, and Hhubs are depicted by blue, red, and green colored dots, respectively; Module Membership (MM) *vs.* Gene Significance (GS) plot for the identification of intramodular high GS (HGS) genes for the trait of interest (**b**), where HGS genes, iHubs, and HGS-iHubs are depicted by plum, blue, and black colored dots, respectively. Here iHub stands for nodes with high MM value

correlation and Student's t-test is used to assign a p-value to the module significance. The modules presenting a significant p-value ($p < 0.05$) are further selected for biological functional analysis.

6.4.2 Modular Analysis for Hub Selection

Modules significantly correlated with one or more traits can be deeply evaluated for identifying relevant hubs, i.e., genes presenting high connectivity values related to the network (overall connectivity) – eHubs – and to the module (intramodular connectivity for each gene based on its Pearson's correlation with all other genes in the module) – iHubs, determined by a *k*Total (x-axis) *vs* *k*Within (y-axis) scatterplot. Genes presenting high *k*Total and *k*Within values are here named HHubs (Fig. 6.5a).

6.4.3 Identification of HGS Genes

Modules showing high correlation with one or more clinical traits are then selected for the identification of genes presenting high GS values (HGS genes), determined by a MM (x-axis) *vs* GS (y-axis) scatterplot. This kind of plot can also reveal iHubs and HGS-iHubs (Fig. 6.5b). Here iHub stands for nodes with high MM value, whereas HGS-iHub stands for nodes with high MM and GS values.

6.4.4 *Functional Enrichment Analysis for Selected Module Genes*

The analysis of co-expression modules involves enrichment analysis – a set of bioinformatics and statistical techniques able to identify classes of molecules (such as genes or proteins) which are over-represented in a large dataset and might have an association with a functional term, a biological pathway, or a disease phenotype (van Dam et al. 2018).

Gene sets of significant trait-associated modules can be submitted to enrichment analyses using, for instance, the Enrichr online web-based tool (Chen et al. 2013; Kuleshov et al. 2016) to identify significantly over-represented terms on GO Biological Process, KEGG pathways, Transcription Factor–PPIs Database, GWAS (Genome-Wide Association Studies), and miRTarBase Database, among other several gene annotations.

6.5 Validation of Transcriptional Networks

The analysis of GCNs based on DNA microarray experiments has multiple applications in life sciences and medicine, ranging from the study of basic cell functions to the identification of disease markers and the molecular mechanisms underlying complex diseases. Therefore, microarray generated data need to be checked for reproducibility and biological significance. Two categories of data validation will be considered here: (i) the technical and biological validation of DNA microarray experiments (Shi et al. 2008); and (ii) the validation of GNCs through interactome analysis (Wang et al. 2014). Additionally, raw microarray data and experimental design should be deposited in at least one data repository supporting MIAME (minimum information about a microarray experiment)-compliance data (Brazma et al. 2001). Two repositories commonly used for this purpose are: Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) at the National Center for Biotechnology Information, and ArrayExpress – functional genomics data, at the European Bioinformatics Institute (www.ebi.ac.uk/arrayexpress/).

6.5.1 *DNA Microarray Technical Validation*

Good laboratory proficiency and appropriate data analysis are essential to avoid artifactual gene profiles generated from DNA microarrays experiments (Shi et al. 2008). Nevertheless, it is mandatory to check for results erroneously representing either under- or overexpression of specific genes. There are several methods to quantify gene expression using RNA or gene-specific protein detection, such as quantitative real-time PCR (qPCR) and immunohistochemistry, respectively (True and Feng 2005).

A commonly used strategy for microarray technical validation using qPCR is to select in the gene data set those presenting the largest fold changes (statistically significant differentially expressed genes between groups). Here one can use the RNA aliquots from the same biological samples tested in the microarray experiment (Miron et al. 2006). In order to accomplish biological validations, it is necessary to test additional biological samples (not those used in the experiment). This is critical, for instance, for validating certain genes as disease biomarkers (LaPointe et al. 2012). This kind of validation usually encompasses, whenever possible, immunohistochemistry validation (Kujawa et al. 2020).

6.5.2 Interactome Validation of GCNs and WGCNA: Tools for Gene Function Discovery

Interactome analysis, particularly protein–protein interaction (PPI) networks (where nodes stand for proteins and edges for the physical interactions), have been used in many different areas, from the study of protein function to disease prognosis (Taylor et al. 2009), being a very useful tool for disease-gene identification (del Rio et al. 2009; Barabási et al. 2011, Carter et al. 2013; Wang et al. 2014). This kind of analysis also allows the in silico validation of GNC data. Protein–protein interaction (PPI) networks for GCN validation may be constructed using proteins corresponding to each of the selected hubs, VIPs, and high-hubs of a particular GCN (Bando et al. 2013; Moreira-Filho et al. 2015). Several major primary protein databases are available for PPI networks, such as APID, BIND, BioGRID, DIP, HPRD, IntAct, and MINT (De Las Rivas and Fontanillo 2010; Khatun et al. 2020; Armingol et al. 2021). Data analysis and visualization are accomplished through Cytoscape. Figure 6.2g, h shows the patient and the control groups' DE GCNs used as a demonstrative example along this chapter. Essentially, the software helps to search for interactions among the selected GCN genes (i.e., their corresponding proteins) and their neighbors in the human interactome. Considering our patient versus control example, these neighbors could participate in some disease-related metabolic pathways, thus indicating that the selected GCN genes are involved in the molecular mechanism of that disease.

The integrative analysis of GCN and PPI data has proven to be very helpful for disclosing changes in steady states that characterize the transitions between health and disease (Sahni et al. 2013, 2015), and for finding common genomic drivers beyond apparently distinct pathophenotypes (Cristino et al. 2014; Nangraj et al. 2020). This approach is also advantageous for identifying disease subtypes. For instance, through GCN and interactome analysis of hippocampal CA3 surgical explants our group was able to reveal pathogenic and compensatory pathways in febrile and afebrile refractory mesial temporal lobe epilepsy (RMTLE), as well as distinct molecular pathways for early- and late-onset RTLME with childhood febrile seizures (Bando et al. 2013; Moreira-Filho et al. 2015). Recently, we employed WGCNA for integrating clinical, histopathological (dentate gyrus), and transcriptomic (CA3) data from a cohort of RMTLE patients and found transcriptional modules highly correlated with age of disease onset, cognitive dysfunctions, and granule cell alterations. We also found 15 genes with high gene significance values, which have the potential to be novel biomarkers and/or therapeutical targets (Bando et al. 2021). The application of WGCNA for analyzing proteomic and metabolomic datasets (Pei et al. 2017) and the integration of PPI and WGCNA computational tools (Nangraj et al. 2020) opened new perspectives for the retrieval of shared and distinct hub signatures underlying pathophenotypes. Now, with the continued reduction in costs and processing time, computer scientists and life scientists are struggling to integrate all omics, what implies not only to deal with very large datasets, but, and fundamentally, to tackle the hard tasks of normalization, data dimensionality reduction, statistical validation, data storage, etc. (Manzoni et al. 2018; Misra et al. 2018; Turek et al. 2020; Vlachavas 2021). Nevertheless, the rewards seem tempting: omics integration is essential for translational research in the era of precision medicine.

Acknowledgments This work was funded by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) research grant 2015/22308-2 and by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grant 3068936/2018-5.

References

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118:4947–4957

Albert R, Jeong H, Barabási AL (2008) Error and attack tolerance of complex networks. Nature 406:378–382

Allen KD, Coffman CJ, Golightly YM et al (2010) Comparison of pain measures among patients with osteoarthritis. J Pain 11:522–527

Ang JC, Mirzal A, Haron H et al (2016) Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinform 13:971–989

Armingol E, Officer A, Harismendy O et al (2021) Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet 22:71–88

Arunkumar C, Sooraj MP, Ramakrishnan S (2017) A comparative performance evaluation of supervised feature selection algorithms on microarray datasets. Procedia Comput Sci 115:209–217

Azevedo H, Moreira-Filho CA (2015) Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma. Sci Rep 5:16830

Bando SY, Silva FN, Costa L d F et al (2013) Complex network analysis of CA3 transcriptome reveals pathogenic and compensatory pathways in refractory temporal lobe epilepsy. PLoS One 8:e79913

Bando SY, Iamashita P, Guth BE et al (2017) A hemolytic-uremic syndrome-associated strain O113:H21 Shiga toxin-producing Escherichia coli specifically expresses a transcriptional module containing dicA and is related to gene network dysregulation in caco-2 cells. PLoS One 12(12):e0189613

Bando SY, Iamashita P, Silva FN et al (2019) Dynamic gene network analysis of caco-2 cell response to Shiga toxin-producing *Escherichia coli*-associated hemolytic-uremic syndrome. Microorganisms 7(7):195

Bando SY, Bertonha FB, Pimentel-Silva LR et al (2021) Hippocampal CA3 transcriptional modules associated with granule cell alterations and cognitive impairment in refractory mesial temporal lobe epilepsy patients. Sci Rep 11(1):10257

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113

Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 13:56–68

Benson M, Breitling R (2006) Network theory to understand microarray studies of complex diseases. Curr Mol Med 6:695–701

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25:163–177

Brazma A, Hingcamp P, Quackenbush J et al (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29:365–371

Bumgarner R (2013) Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol Chapter 22:Unit 22.1

Cai JJ, Borenstein E, Petrov DA (2010) Broker genes in human disease. Genome Biol Evol 2:815–825

Carter H, Hofree M, Ideker T (2013) Genotype to phenotype via network analysis. Curr Opin Genet Dev 23:611–621

Chaussabel D, Baldwin N (2014) Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol 14:271–280

Chen EY, Tan CM, Kou Y et al (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14:128

Cho D-Y, Kim Y-A, Przytycka TM (2012) Chapter 5: Network biology approach to complex diseases. PLoS Comput Biol 8(12):e1002820

Clauset A, Shallizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51:661–703

Cockrum C, Kaneshiro KR, Rechtsteiner A et al (2020) A primer for generating and using transcriptome data and gene sets. Development 147(24):dev193854

Costa L d F, MAR T, Silva FN (2008) Concentric characterization and classification of complex network nodes: application to an institutional collaboration network. Physica A 387:6201–6214

Costa L d F, Oliveira ON Jr, Travieso G et al (2011) Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Adv Phys 60:329–412

Costanzo M, Baryshnikova A, Bellay J et al (2010) The genetic landscape of a cell. Science 327:425–431

Costa-Silva J, Domingues D, Lopes FM (2017) RNA-Seq differential expression analysis: an extended review and a software tool. PLoS One 12(12):e0190152

Cristino AS, Williams SM, Hawi Z et al (2014) Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. Mol Psychiatry 19:294–301

De Las Rivas J, Fontanillo C (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol 6:e1000807

Del Rio G, Koschutzki D, Coello G (2009) How to identify essential genes from molecular networks? BMC Syst Biol 3:102

Elo LL, Järvenpää H, Oresic M et al (2007) Systematic construction of gene co-expression networks with applications to human T helper cell differentiation process. Bioinformatics 23:2096–2103

Faro A, Giordano D, Spampinato C (2012) Combining literature text mining with microarray data: advances for system biology modeling. Brief Bioinform 13:61–82

Flake GW, Lawrence SR, Giles CL et al (2002) Self-organization and identification of web communities. IEEE Comput 35:66–71

Freeman LC (1978) Centrality in social networks: conceptual clarification. Soc Networks 1:215–239

Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement software. Pract Exp 21:1129–1164

Gaiteri C, Ding Y, French B et al (2014) Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav 13:13–24

Galán-Vásquez E, Perez-Rueda E (2019) Identification of modules with similar gene regulation and metabolic functions based on co-expression data. Front Mol Biosci 6:139

Geraci F, Saha I, Bianchini M (eds) (2020) RNA-seq analysis: methods, applications and challenges. Frontiers Media SA, Lausanne

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci U S A 99:7821–7826

Gysi DM, Nowick K (2020) Construction, comparison and evolution of networks in life sciences and other disciplines. J R Soc Interface 17(166):20190610

Horvath S, Dong J (2008) Geometric interpretation of gene co-expression network analysis. PLoS Comput Biol 4:e1000117

Hu JX, Thomas CE, Brunak S (2016) Network biology concepts in complex disease comorbidities. Nat Rev Genet 17:615–629

Ishiwata RR, Morioka MS, Ogishima S et al (2009) BioCichlid: central dogma-based 3D visualization system of time-course microarray data on a hierarchical biological network. Bioinformatics 25:543–544

Joshi A, Rienks M, Theofilatos K et al (2021) Systems biology in cardiovascular disease: a multiomics approach. Nat Rev Cardiol 18:313–330

Kakati T, Bhattacharyya DK, Barah P et al (2019) Comparison of methods for differential co-expression analysis for disease biomarker prediction. Comput Biol Med 113:103380

Khatun MS, Shoombuatong W, Hasan MM et al (2020) Evolution of sequence-based bioinformatics tools for protein-protein interaction prediction. Curr Genomics 21:454–463

Kujawa KA, Zembala-Nożyńska E, Cortez AJ et al (2020) Fibronectin and periostin as prognostic markers in ovarian cancer. Cell 9:149

Kuleshov MV, Jones MR, Rouillard AD et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44(W1):W90–W97

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559

Langfelder P, Mischel PS, Horvath S (2013) When is hub gene selection better than standard meta-analysis? PLoS One 8:e61505

LaPointe LC, Pedersen SK, Dunne R et al (2012) Discovery and validation of molecular biomarkers for colorectal adenomas and cancer with application to blood testing. PLoS One 7(1):e29059

Lee WP, Tzou WS (2009) Computational methods for discovering gene networks from expression data. Brief Bioinform 10:408–423

Li A, Horwath S (2009) Network module detection: affinity search technique with the multi-node topological overlap measure. BMC Res Notes 2:142

Liu YY, Slotine JJ, Barabási AL (2011) Controllability of complex networks. Nature 473:167–173

Liu R, Li M, Liu ZP et al (2012a) Identifying critical transitions and their leading biomolecular networks in complex diseases. Sci Rep 2:813

Liu YY, Slotine JJ, Barabási AL (2012b) Control centrality and hierarchical structure in complex networks. PLoS One 7(9):e44459

Mahendran N, Durai R, Vincent PM et al (2020) Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. Front Genet 11:603808

Mantione KJ, Kream RM, Kuzelova H et al (2014) Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res 20:138–142

Manzoni C, Kia DA, Vandrovcova J et al (2018) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform 19:286–302

Masuda N, Konno N (2006) VIP-club phenomenon: emergence of elites and masterminds in social networks. Soc Networks 28:297–309

Mcauley JJ, Costa L d F, Caetano TS (2007) Rich-club phenomenon across complex network hierarchies. Appl Phys Lett 91:084103

Milo R, Shen-Orr S, Itzkovitz S et al (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827

Miron M, Woody OZ, Marcil A et al (2006) A methodology for global validation of microarray experiments. BMC Bioinformatics 7:333

Misra BB, Langefeld CD, Olivier M et al (2018) Integrated omics: tools, advances, and future approaches. J Mol Endocrinol 13:JME-18-0055

Moreira-Filho CA, Bando SY, Bertonha FB et al (2015) Community structure analysis of transcriptional networks reveals distinct molecular pathways for early- and late-onset temporal lobe epilepsy with childhood febrile seizures. PLoS One 10(5):e0128174

Moreira-Filho CA, Bando SY, Bertonha FB et al (2016) Modular transcriptional repertoire and MicroRNA target analyses characterize genomic dysregulation in the thymus of Down syndrome infants. Oncotarget 7:7497–7533

Nangraj AS, Selvaraj G, Kaliamurthi S et al (2020) Integrated PPI- and WGCNA-retrieval of hub gene signatures shared between Barrett’s esophagus and esophageal adenocarcinoma. Front Pharmacol 11:881

Newman MEJ (2006) Modularity and community structure in networks. PNAS 103:8577–8582

Newman MEJ (2010) Networks: an introduction. Oxford University Press, New York

Pavlopoulos GA, O’Donoghue SI, Satagopam VP et al (2008) Arena3D: visualization of biological networks in 3D. BMC Syst Biol 2:104. http://www.biomedcentral.com/1752-0509/2/104

Pei G, Chen L, Zhang W (2017) WGCNA application to proteomic and metabolomic data analysis. Methods Enzymol 585:135–158

Prifti E, Zucker JD, Clement K et al (2008) Funnet: an integrative tool for exploring transcriptional interactions. Bioinformatics 24:2636–2638

R Core Team (2012) R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. http://www.R-project.org/

Rao MS, Van Vleet TR, Ciurlionis R et al (2019) Comparison of RNA-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. Front Genet 9:636

Ravasz E, Somera AL, Mongru DA (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555

Rosenkrantz JT, Aarts H, Abee T et al (2013) Non-essential genes form the hubs of genome scale protein function and environmental gene expression networks in Salmonella enterica serovar Typhimurium. BMC Microbiol 13:294

Saeed A, Sharov V, White J et al (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34:374–378

Sahni N, Yi S, Zhong Q et al (2013) Edgotype: a fundamental link between genotype and phenotype. Curr Opin Genet Dev 23:649–657

Sahni N, Yi S, Taipale M et al (2015) Widespread macromolecular interaction perturbations in human genetic disorders. Cell 161:647–660

Saito R, Smoot ME, Ono K et al (2012) A travel guide to cytoscape plugins. Nat Methods 9:1069–1076

Schroeder A, Mueller O, Stocker S et al (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7:3

Shen-Orr SS, Milo R, Mangan S et al (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31:64–68

Shi L, Perkins RG, Fang H et al (2008) Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. Curr Opin Biotechnol 19:10–18

Sieberts SK, Schadt EE (2007) Moving toward a system genetics view of disease. Mamm Genome 18:389–401

Silva FN, Rodrigues FA, Oliveira Junior ON et al (2013) Quantifying the interdisciplinarity of scientific journals and fields. J Informetr 7:469–477

Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model-based indices. BMC Bioinformatics 13:328

Taylor IW, Linding R, Wade-Farley D et al (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol 27:199–204

True L, Feng Z (2005) Immunohistochemical validation of expression microarray results. J Mol Diagn 7:149–151

Tuck DP, Kluger HM, Kluger Y (2006) Characterizing disease states from topological properties of transcriptional regulatory networks. BMC Bioinformatics 7:236

Turek C, Wrobel S, Piwowar M (2020) OmicsON – integration of omics data with molecular networks and statistical procedures. PLoS One 15(7):e0235398

van Dam S, Võsa U, van der Graaf A et al (2018) Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform 19:575–592

Vermeulen R, Schymanski EL, Barabási AL et al (2021) The exposome and health: Where chemistry meets biology. Science 367:392–396

Villa-Vialaneix N, Liaubet L, Laurent T et al (2013) The structure of a gene co-expression network reveals biological functions underlying eQTLs. PLoS One 8:e60045

Vlachavas EI, Bohn J, Ückert F et al (2021) Detailed catalogue of multi-omics methodologies for identification of putative biomarkers and causal molecular networks in translational cancer research. Int J Mol Sci 22:2822

Wang H, Zheng H (2012) Correlation of genetic features with dynamic modularity in the yeast interactome: a view from the structural perspective. IEEE Trans Nanobiosci 11:244–250

Wang Q, Tang B, Song L et al (2013) 3DScapeCS: application of 3 dimensional, parallel, dynamic network visualization in Cytoscape. BMC Bioinformatics 14:322. http://www.biomedcentral.com/1471-2105/14/322

Wang XD, Huang JL, Yang L et al (2014) Identification of human disease genes from interactome network using graphlet interaction. PLoS One 9:e86142

Watkinson J, Liang KC, Wang X et al (2009) Inference of regulatory gene interactions from expression data using three-way mutual information. Ann N Y Acad Sci 1158:302–313

Watts DJ, Strogatz SH (1998) Collective dynamics of “small word” networks. Nature 393:440–442

Weirauch MT (2011) Gene expression network for the analysis of cDNA microarray data. In: Dehmer M, Emmert-Streib F, Graber A, Salvador A (eds) Applied statistics for network biology: methods in systems biology, vol 1. Weinheim, Wiley-Blackwell, pp 215–250

Weiss JM, Karma A, Robb MacLellan W et al (2012) “Good enough solutions” and the genetics of complex diseases. Circ Res 111:493–504

Winterbach W, Van Mieghem P, Reinders M et al (2013) Topology of molecular interaction networks. BMC Syst Biol 7:90

Wu X, Wang W, Zheng WX (2012) Inferring topologies of complex networks with hidden variables. Phys Rev E 86:046106

Wu L, Li M, Wang JX et al (2019) Controllability and its applications to biological networks. J Comput Sci Technol 34:16–34

Yu H, Kim PM, Sprecher E et al (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3:e59

Yuan Z, Zhao C, Di Z et al (2013) Exact controllability of complex networks. Nat Commun 4:2447

Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:Article17

Zhang J, Ji Y, Zhang L (2007) Extracting three-way gene interactions from microarray data. Bioinformatics 23:2903–2909

Zhao W, Langfelder P, Fuller T et al (2010) Weighted gene coexpression network analysis: state of the art. J Biopharm Stat 20:281–300

Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21:1010–1024

Chapter 7
Comparative Analysis of Packages and Algorithms for the Analysis of Spatially Resolved Transcriptomics Data

Natalie Charitakis, Mirana Ramialison, and Hieu T. Nim

7.1 Introduction

Despite the natural stochasticity that can disrupt biological processes such as organ development, biological systems consistently produce the same gene expression pattern with sufficient robustness such that the embryo forms correctly (nearly) every time. Furthermore, the genes typically work together in networks, requiring a systems-wide transcriptomic approach to fully understand the spatial expression patterns. Many of these create well-defined regions of cells within developing tissues that can be easily reproduced, demonstrating how the spatial location of the gene regulatory networks is critical for the proper formation of tissues (Exelby et al. 2021). Determining these networks is an active study area in the emerging field of 'spatial biology', and calls for specialised computational techniques, many of which have been developed very recently.

The merits and limitations of single-cell RNA Sequencing (scRNA-Seq) have been well established (Hwang et al. 2018; Chen et al. 2019) and the method

Co-senior authors: Mirana Ramialison and Hieu T. Nim.

N. Charitakis
Murdoch Children's Research Institute, Parkville, VIC, Australia

Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia
e-mail: natalie.charitakis@mcri.edu.au

M. Ramialison (✉) · H. T. Nim (✉)
Murdoch Children's Research Institute, Parkville, VIC, Australia

Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia

Australian Regenerative Medicine Institute and Systems Biology Institute Australia, Monash University, Clayton, VIC, Australia
e-mail: mirana.ramialison@mcri.edu.au; hieu.nim@mcri.edu.au

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_7

successfully applied across varying organs and conditions (Karaayvaz et al. 2018; Regev et al. 2017; Dong et al. 2018; He et al. 2020; Ximerakis et al. 2019; Tiklová et al. 2019; Zhou et al. 2021). scRNA-Seq is capable of identifying rare cell populations, including in disease states and developmental stages; however, the method yields noisy, variable data with lots of technical variation (Chen et al. 2019). Despite scRNA-Seq allowing for the study of cellular heterogeneity and cell type hierarchy, the loss of spatial information prevents the systematic study of physiological structure/function relationships in various tissues and organs. This was part of the drive in the development of spatial transcriptomics (ST) (Marx 2021) (now commercialised by 10x Genomics under the name Visium) and other spatially resolved transcriptomics (SRT) methods. The spatially resolved gene expression pattern within the context of a tissue is critical to achieving a full understanding of disease states and tissue development and function and the ability to investigate this is achievable using SRT (Ståhl et al. 2016).

Spatial transcriptomics is an area that is becoming more widely used and will continue to expand in the upcoming years (Marx 2021). Having been featured as Nature's 'Method of the Year' in 2020, the technology and the analytical opportunities it provides are going to keep growing rapidly (Marx 2021). As demonstrated in Fig. 7.1, the number of papers published on spatial transcriptomics has greatly increased since 2016, when the first technology named 'spatial transcriptomics' was published (Ståhl et al. 2016). Offering unprecedented spatial context to transcriptomic data presents an invaluable tool for studying tissues and their cellular

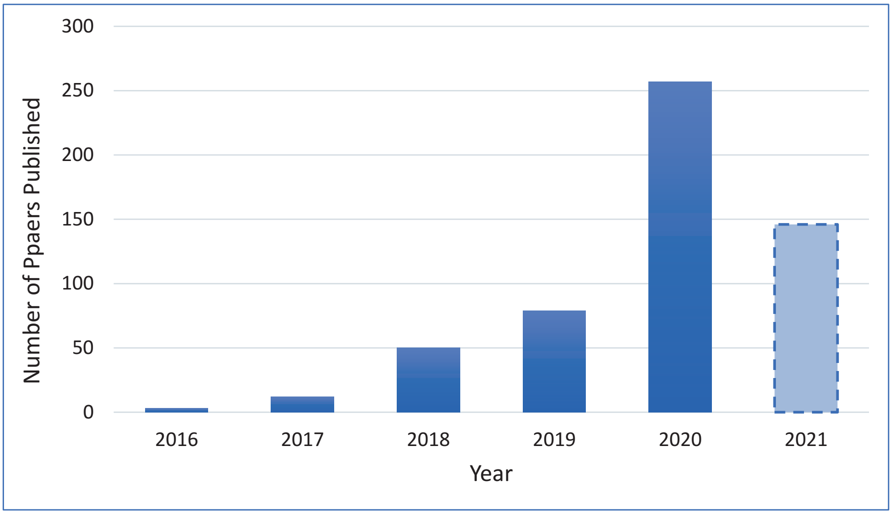


Fig. 7.1 Number of papers returned when a search was performed using the keywords 'Spatial Transcriptomics' using the software 'Publish or Perish'(Harzing 2016) to search PubMed and to manually search bioRvix, with the additional parameter of papers published from 01/01/2016 to 16/04/2021. Papers identified by searching both databases were consolidated; note that this is not a comprehensive view of all papers published on the topic since 2016. Bars in light blue with a dotted outline indicate that not all papers for the calendar year have been included

composition. As early as 2017, the merits of applying SRT to the discovery of spatial organisation of gene expression to improve transcriptional classification of cell types and localisation within a tissue had been discussed and even put to the test (Lein et al. 2017; Shah et al. 2016). The potential applications of this technology are continuously improving and expanding, as demonstrated by the integration of different methods to improve the resolution of current SRT methods (Moncada et al. 2020). The different techniques available to generate SRT data and their merits have been discussed (Lein et al. 2017; Crosetto et al. 2015; Asp et al. 2020; Waylen et al. 2020), but a review of data analysis tools is as of yet lacking. With an emphasis on obtaining spatially resolved data sets with single-cell resolution (Marx 2021), the method, aims and approaches to integrate and analyse the data generated are still in flux, with a clear 'gold standard' yet to distinguish itself. This chapter discusses some of the current packages and pipelines available to perform this analysis (Table 7.1).

7.2 Methods for Downstream Analysis of Spatially Resolved Transcriptomics Data

As identifying the spatial expression patterns of genes and how they vary across a tissue is a critical aim of spatial transcriptomics, many purpose-built tools for analysis of this data aim to identify spatially variable genes (SVGs) (Box 7.1) (Exelby et al. 2021). Building on the concept of highly variable genes in scRNA-Seq analysis, SVGs have a pattern of expression that depends on their location in the tissue and can give insight into biological function (Svensson et al. 2018). A complication of analysing these spatial transcriptomics data sets is accurately accounting for the spatial correlation across samples (Li et al. 2021), and different methods can be employed to tackle this problem. Various packages have been developed in primarily R or Python and are currently available to identify SVGs in spatial transcriptomic data sets.

Box 7.1

One key aim of analysing RNA-Seq and scRNA-Seq datasets is to identify differentially expressed genes (DEGs) between two groups from within a group of highly variable genes (HVGs). DEGs are identified between two groups when a gene's expression is statistically significantly different between the two groups present (Exelby et al. 2021). While this approach has yielded many important findings, it removes organisational context from the groups in question, something that can be recovered using spatial transcriptomics (Marx 2021). This new technology has shifted the goalposts for transcriptomics analysis, resulting in many bioinformatics packages dedicated to discovering *spatially variable genes* (SVGs) (Svensson et al. 2018; Li et al. 2021; Sun et al. 2019; Edsgärd et al. 2018; Hao et al. 2021; Zhang et al. 2018). As the name suggests, these genes will have amplified expression in certain regions of the tissue or sample, often displaying an underlying pattern (Svensson et al. 2018; Hu et al. 2020). Determining the best method to achieve the most biologically accurate results and computational efficiency is challenging, and research in this area is ongoing.

Table 7.1 Comparison of computational packages for analysis of spatial transcriptomics data sets

Purpose	Package name	Main method	Implementation	Pros	Cons	GitHub
Identifying SVGs	SpatialDE	GP Regression	Python	Currently most popular package in this category	Labels genes as SVGs that have very low expression and two normalisation steps	https://github.com/Teichlab/SpatialDE
	SPARK	Generalised Linear Spatial Model	R	Does not require data to be normalised and controls for type I error	Accuracy not significant improvement on SpatialDE	https://github.com/xzhoulab/SPARK
	Trendsceek	Marked Point Process	R	Low false positives reported	Identifies very low number of SVGs and ineffective on larger data sets	https://github.com/edsgard/trendsceek
	BOOST-GP	Bayesian Hierarchical Model	R	Accuracy rate is better than other packages in data sets with many 0 counts	Accuracy rate still low in the presence of many 0 counts	https://github.com/Minzhe/BOOST-GP
	SOMDE	SOM	Python	Able to efficiently identify SVGs even in very large data sets	In low dropout rate datasets not as good as SpatialDE	https://github.com/WhirlFirst/somde
	scGCO	Graph Cut Algorithm	Python	Results were more reproducible than SpatialDE. Can be used on data sets with millions of cells	~35% of labelled SVGs not reproducible	https://github.com/WangPeng-Lab/scGCO

Identifying SVGs + other capabilities	GP counts	GP Regression	Python	Can determine temporal trajectories and perform pseudotime analysis	Efficiency on larger data sets unclear	https://github.com/ManchesterBioinference/GPcounts
	STUtility	Spatial Autocorrelation	R	Image processing and the ability to create a 3D model from multiple samples	Accuracy in identifying SVGs and defining tissue heterogeneity not comprehensively reviewed	https://github.com/jbergenstrahle/STUtility
Assigning lost transcripts	Sparcle	MLE	Python	Unique capability and can be used alongside other packages	Developed specifically for smFISH	https://github.com/sandhya212/Sparcle_for_spot_reassignments
Cell type identification	SpatialDWLS	DWLS	R	*A priori* knowledge can be incorporated	Performance not validated to other packages on real datasets	https://github.com/rdong08/spatialDWLS_dataset
	FICT	Generative Mixture Model	Python	Addresses problem of over-reliance on expression data	Performance drops on data sets with less cells	https://github.com/haotianteng/FICT
	RCTD	Supervised Learning	R	Can normalise across platform effects of scRNA-Seq and SRT data sets	Requires well-annotated scRNA-Seq data sets	https://github.com/dmcable/RCTD

(continued)

Table 7.1 (continued)

Purpose	Package name	Main method	Implementation	Pros	Cons	GitHub
Spot-to-spot clustering	SpatialCPie	Different Clustering Algorithms	R	Can perform clustering at different resolutions for different subtypes of tissue; cluster graph is a novel method of visualising cluster origin in SRT	Validation against other packages lacking	https://github.com/jbergenstrahle/SpatialCPie
Pipeline	Giotto	–	R	Choice of algorithm for identifying marker genes in cell types, dedicated pipelines for lower resolution SRT data	Validation against different biological tissues collected on different platforms lacking	https://github.com/RubD/Giotto
	Squidpy	–	Python	Modular, so can incorporate other packages in analysis	Cellular neighbourhoods not very reproducible	https://github.com/theislab/squidpy

7.2.1 Identifying Spatially Variable Genes

Among these, SpatialDE is a popular package based on Gaussian process (GP) regression, which can clearly identify localized gene expression patterns for data sets containing temporal and/or spatial annotations (Svensson et al. 2018). SpatialDE can recognise SVGs by creating a model with two different terms reflecting distinct variance present in the data set. The first term captures the non-spatial variance present within the data, while the second aims to capture the spatially related variance of gene expression within the data set, with the assumption that the covariance between a cell's gene expression profile decreases with an increase in distance between the cells (Svensson et al. 2018). A ratio calculated using these terms can then be used as a measure of the level of gene expression variance attributable to spatial location (Svensson et al. 2018). These are the key parameters used to fit the Gaussian model in a computationally efficient manner (Svensson et al. 2018). Testing to prove whether statistically significant SVGs are present is performed by comparing this model to a second one which lacks the spatial covariance parameter that represents a data set in which spatial localisation has no effect on gene expression patterns (Svensson et al. 2018). This process is repeated for each gene, and after correcting for multiple testing, the SVGs can be pulled out of the data set (Svensson et al. 2018). SpatialDE has the capability of taking this a step further by creating models with different covariance functions for SVGs and comparing them, this is in addition to the initial 10 Gaussian kernels it tests before selecting that with the lowest p-value. This creates the ability to determine whether each SVGs is most accurately expressed as a linear, periodic or general expression model (Svensson et al. 2018). However, for the data to fit certain underlying assumptions of the Gaussian model, two normalisation steps are performed, the first being a variance stabilising transformation (Svensson et al. 2018; Sun et al. 2019). It may affect the package's performance as the assumptions underlying the model and the necessary data transformations do not truly reflect the nature of the data (Li et al. 2021). A further functionality of SpatialDE is that it can implement an unsupervised learning technique built on the Gaussian Mixture Model to apply automatic expression histology (AEH), which can group together SVGs by their spatial expression pattern using hidden patterns learnt from the data (Svensson et al. 2018). The observation that SpatialDE may introduce false positives by labelling genes with low levels of expression as SVGs is an area which requires further investigation and can be improved upon in future releases of the package (Sun et al. 2019).

A package with the same goal as SpatialDE is SPARK (Spatial Pattern Recognition via Kernels), which employs a generalised linear spatial model (GLSM) with different spatial kernels to identify SVGs (Sun et al. 2019). This model was built on previous work to take into consideration the effects of spatial correlation and covariate measurement error; it was built and tested on 2D data; however, it is capable of being expanded to 3D data sets (Sun et al. 2019). As in the case of SpatialDE, SPARK models gene expression for each gene across all spatial coordinates; however, this model operates under the assumption that the spatial data is non-Gaussian (Sun et al. 2019). SPARK builds on other GLSMs by developing a hypothesis testing framework for the model (Sun et al. 2019). The power of this hypothesis testing is linked to how the spatial kernel function accurately represents the spatial pattern of the gene represented in the model; and as different gene expression patterns will most accurately be represented by different spatial kernel functions, SPARK considers 10 different kernels (similarly to SpatialDE) based on commonly observed biological patterns (Sun et al. 2019). Due to the heuristic nature of these kernels, this process could introduce biases that lead that package to choose more commonly observed biological patterns. SPARK can work with large data sets as it employs a penalised quasi-likelihood (PQL) algorithm for parameter estimation to circumvent the problem of the difficulty in solving GLSMs in short periods of time; this algorithm informs the parameters used in each of the spatial kernel functions. It further improves on the packages available at the time of publication, SpatialDE and Trendsceek, by not performing a normalisation step on the data, which decreases the power of the analysis (Sun et al. 2019). A drawback of SpatialDE that SPARK corrects for is to control for type 1 errors through the Cauchy combination rule, thus giving it additional power when identifying SVGs (Sun et al. 2019). The Cauchy combination rule groups the p-values generated from each spatial kernel function into a single p-value while still controlling for type 1 errors, which results in a single p-value per gene (Sun et al. 2019). The final steps involve controlling for FDR across all p-values and then determining which are SVGs (Sun et al. 2019). While SpatialDE and SPARK share the use of parametric test statistics, there are a few critical differences between the packages (Sun et al. 2019). As previously mentioned, SPARK does not model normalised data, while SpatialDE can only approximate p-values; SpatialDE first calculates an exact p-value per gene; and once it obtains the initial set of statistically significant genes, SpatialDE then performs additional analysis to determine their p-values (Sun et al. 2019). Furthermore, when validated against multiple data sets, it performed just as well or better than SpatialDE and Trendsceek (described in next paragraph) (Sun et al. 2019). When its ability to calculate true positives in two simulated data sets was tested across a total of six different spatial expression patterns with varying FDRs, SPARK outperformed Trendsceek and had better results than SpatialDE (Sun et al. 2019). While with certain simulated data sets, SPARK and Trendsceek performed similarly in computing well-calibrated p-values, but SpatialDE did not identify certain SVGs present (Sun et al. 2019). While the SPARK paper only tests the package's performance against SpatialDE and Trendsceek, it outperformed both in terms of the number of SVGs identified when validated against a spatial transcriptomics mouse olfactory bulb data set (Sun et al. 2019). However, not all genes identified by SpatialDE overlapped with those identified by SPARK (Sun et al. 2019). Despite this, the newly identified SVGs are in line with markers specific to the tissue they were annotated in, and GO enrichment analysis adds further confidence that the majority of these newly identified SVGs are biologically relevant (Sun et al. 2019). In terms of computational efficiency, when running with 10 parallel CPU threads, SPARK was more computationally efficient than the same analysis run on a single-threaded SpatialDE (although the difference in this instance is minimal) and

Trendsceek; its single-threaded performance is consistently less efficient than SpatialDE across 4 datasets of varying sizes (Sun et al. 2019).

Trendsceek is one of the earlier packages developed to identify SVGs using a non-parametric approach (Edsgärd et al. 2018). Trendsceek individually assesses each gene and normalises its expression through a log10 transformation (Edsgärd et al. 2018). It relies on a marked point process to model gene expression and cell location and later will test the null hypothesis by generating four non-parametric test statistics (Edsgärd et al. 2018). These four test statistics yield four p-values and a gene with a minimum of 1 p-value ≤ 0.05 after adjustment for multiple testing using the Benjamini-Hochberg method is determined to be an SVG (Edsgärd et al. 2018). A key difference that separates Trendsceek from SpatialDE and SPARK is its computing of non-parametric test statistics, meaning it lacks an underlying generative model. Trendsceek was tested against simulated data sets, and it demonstrated very low power to identify SVGs when they were present if less than 5% of cells in the data set had varying levels of expression (Edsgärd et al. 2018). This implies that as SRT datasets continue to increase in size, Trendsceek will not be able to distinguish SVGs present in a very small subset of cells within a tissue. When Trendsceek's performance in identifying SVGs across two spatial transcriptomics data sets is compared to SpatialDE and SPARK, it identified fewer SVGs, with numbers almost 10 times lower than the other packages (Sun et al. 2019). When compared to different packages in other studies, Trendsceek struggled to identify SVGs in real datasets, while other packages were able to (Sun et al. 2019).

Each new package developed aims to address the shortcomings of those already published; for example, BOOST-GP claims that many popular substitutes such as SpatialDE, SPARK and Trendsceek do not account for the substantial proportion of zero counts present in the data set and the effect the sparsity of the data can have analysis (Li et al. 2021). Therefore, BOOST-GP puts forth a new Bayesian hierarchical model aimed at accounting for the considerable number of zero counts present in spatial data sets, that other packages published up to this point had neglected (Li et al. 2021). A key difference to other packages is that BOOST-GP employs a negative binomial distribution when modelling count data, which should account for its observed over-dispersion (Li et al. 2021). This resembles the methods used by popular bulk RNA-Seq analysis packages rather than other spatial transcriptomics packages explored thus far (Li et al. 2021). BOOST-GP's performance was compared to that of SpatialDE's, SPARK and Trendsceek when there were false zeros present in the data, and BOOST-GP was clearly most adept at handling this complication, even if it still presented significant difficulties in retrieving a good Matthews correlation coefficient (used to determine the tool's accuracy) on a synthetic data set (Li et al. 2021). Furthermore, depending on the spatial pattern of the expression of the gene, the accuracy of BOOST-GP can differ slightly (Li et al. 2021). Alternatively, when the tool was tested on two real data sets, it was found that SPARK identified more SVGs than BOOST-GP; however, SpatialDE discovered the least (Li et al. 2021). In the analysis of human breast cancer data, despite identifying fewer SVGs than SPARK, BOOST-GP was able to identify novel, biologically relevant terms in the GO analysis, adding to its value in the analysis of SRT data (Li et al. 2021).

As larger datasets become increasingly common, packages must be created to efficiently analyse the vast amounts of data generated by SRT experiments. One of the newer packages is SOMDE (Hao et al. 2021). Built-in python, SOMDE aims to identify SVGs in large-scale datasets (Hao et al. 2021). By using a self-organising map (SOM) neural network and a Gaussian process to model the data, it can identify SVGs in large datasets much faster than SpatialDE, SPARK or Trendsceek (Hao et al. 2021). This is achieved as the data is organised into different nodes by the SOM neural network, the Gaussian process is used at the level of the nodes to identify the SVGs present in the data (Hao et al. 2021). The organisation of data into nodes minimises the sample space while preserving the original spatial organisation and expression data (Hao et al. 2021). The next stage which uses a Gaussian process identifies the SVGs from the reduced sample space (Hao et al. 2021). As seen in packages such as SpatialDE and BOOST-GP, the Gaussian process is a popular method for identifying SVGs (Svensson et al. 2018; Li et al. 2021). SOMDE also uses a log ratio test similar to that employed by SpatialDE to test the statistical significance of the spatial expression variability of each gene (He et al. 2020). When SOMDE was applied to discover the SVGs of five different data sets, it was able to do so without significant increase in computational time as the size of the data set increased, yielding results in under 5 min for the largest data set with over 20,000 data sites (Hao et al. 2021). It also demonstrated a faster running time compared to Giotto and SpatialDE on three differently sized data sets used for validation (Hao et al. 2021). Despite this, the package lacks validation on a data set of single-cell resolution (Hao et al. 2021). When its performance was compared to scGCO and SpatialDE on a simulated data set, SOMDE consistently outperformed scGCO but only had an improved performance compared to SpatialDE when a high dropout rate is incorporated into the data set (Hao et al. 2021). When its performance was compared to real data sets, most of the SVGs identified by SOMDE overlap with those identified by packages like scGCO, SPARK and SpatialDE (Hao et al. 2021).

Other methods have been developed to identify SVGs that differ from those presented thus far. One of these methods has been implemented in a python package called scGCO, which employs graph cut algorithms to identify SVGs (Zhang et al. 2018). scGCO first produces a graph by performing a Delaunay triangulation in which only true cell neighbours are connected by edges, allowing an accurate representation of cellular interactions in a sparse graph which is not memory intensive (Zhang et al. 2018). Subsequently, Voronoi diagrams are created which have previously been used to model cells (Zhang et al. 2018). Using a Markov random field (MRF) model and adapting methods traditionally used in object identification in images, scGCO can classify cells into two categories which provide efficient, low polynomial time computing and a result which is globally optimal (Zhang et al. 2018). Much like SpatialDE, scGCO employs Gaussian Mixture modelling but uses it to classify each gene's expression to ensure more accurate classification of cell types based on their gene expression (Svensson et al. 2018; Zhang et al. 2018). The performance of SVG was tested against a spatial transcriptomics data set obtained from a mouse olfactory bulb and compared to results obtained from the same data by SpatialDE (Zhang et al. 2018). A more comprehensive review of scGCO against different packages would be beneficial to obtain a holistic understanding of its improved performance in SVG detection. scGCO successfully identified over 1,000 additional SVGs compared to SpatialDE, and at an FDR cut-off of 0.01 rather than 0.05 (Zhang et al. 2018). The majority of SVGs identified by scGCO were also identified by SpatialDE, and each formed its own spatial pattern. These results were consistent when validation was repeated across replicate mouse olfactory bulb data (Zhang et al. 2018). However, while scGCO yielded a smaller number of unreproducible SVGs across the different replicate data sets than SpatialDE, ~35% of identified SVGs were still unreproducible (an 11% reduction from SpatialDE) (Zhang et al. 2018). If replicate data sets are available for studies, then this is something that should be investigated further across all packages, resulting in the exclusion of non-reproducible SVGs for a more accurate final subset of SVGs. Additionally, when comparing between regions of the mouse olfactory bulb, scGCO was more adept at identifying SVGs than SpatialDE, while neither method entirely recovered all marker genes reported in the study which published the data set (Zhang et al. 2018). Additional validation was performed using data from breast cancer biopsies, with scGCO having a similar improved performed compared to SpatialDE when employed on the mouse olfactory bulb data set. Furthermore, the SVGs identified by SpatialDE within the breast cancer data set did not maintain consistent clustering pattern (Zhang et al. 2018). scGCO's performance on other spatial transcriptomics data sets was equally as robust (Zhang et al. 2018). scGCO also performed better in terms of computational time and memory required than SpatialDE and Trendsceek when used to analyse a simulated data set with up to a million cells.

7.2.2 Identifying Spatially Variable Genes and More

As evidenced by the packages reviewed so far, GPs are a popular method for analysing spatial transcriptomics data as they can model its spatial dependence. To this end, as new packages are developed, many are built on alternative GP regression models, such as GPcounts (BinTayyash et al. 2020). GPcounts can be used to model either spatial or temporal large-scale scRNA-Seq data through modelling count data using a negative binomial (NB) likelihood (BinTayyash et al. 2020). The NB likelihood model should more accurately capture the distribution of gene expression data compared to Gaussian likelihood model as it accounts for possible heteroscedastic noise and the presence of many zero-counts but requires UMI normalisation to be applied (BinTayyash et al. 2020). Furthermore, GPcounts evaluates its performance across different simulated data sets when it implements different underlying likelihood models to determine under which conditions each yields the best results (BinTayyash et al. 2020). Subsequently, it can be observed that employing an NB likelihood was effective in producing accurately identified SVGs in the package BOOST-GP (Li et al. 2021). However, GPcounts's primary aim is not to identify SVGs, it is also able to identify differentially expressed genes (DEGs), perform pseudotime inference and then identify branching genes and discover temporal trajectories, widening its scope compared to most packages (BinTayyash et al. 2020). The GP model is stochastic and non-parametric, and there is a choice of kernel to find one that most accurately models the data, similarly to the step employed by SpatialDE (Svensson et al. 2018), and this is determined by the Bayesian Inference Criterion (BinTayyash et al. 2020). Using SpatialDE as a benchmark, GPcounts builds on and alters many of the steps implemented by SpatialDE (BinTayyash et al. 2020). This applies from the testing procedures used to determine SVGs and DEGs p-values to the type of normalisation applied to the data (BinTayyash et al. 2020). GPcounts has also implemented the additional step of a built-in check during its kernel function hyperparameter estimation to minimise the problems of getting stuck in a local optimum by restarting the optimisation as this is suspected (BinTayyash et al. 2020). This is so far one of the only optimisation-based methods that has implemented this kind of self-check and could give GPcounts a distinct advantage in the accurate identification of SVGs. An improved assessment of GPcounts performance when detecting DEGs would be to evaluate the package on published data sets in addition to the simulated data (BinTayyash et al. 2020). When evaluated for its identification of SVGs, GPcounts did use a real mouse olfactory bulb data set and compared its performance to SpatialDE, SPARK and Trendsceek (BinTayyash et al. 2020). GPcounts identifies the most SVGs out of any of the packages, with the vast majority of identified SVGs at a 5% FDR overlapping with those identified by SpatialDE and SPARK (BinTayyash et al. 2020). The unique SVGs identified by GPcounts have spatial patterns that match those depicted in the Allen Brain Atlas, indicating a high confidence in these findings (BinTayyash et al. 2020). GPcounts also identified 90% of the biologically important marker genes expressed in the dataset, although SPARK had a similar performance as it identified 80% (BinTayyash et al. 2020), while SpatialDE identified only 30% of the marker genes (BinTayyash et al. 2020).

Certain frameworks have been developed with a particular SRT technology in mind, in combination with addressing an area of data analysis the developers deem lacking. One of these is the STUtility workflow created in R and based and built on the Seurat analysis tool (Bergenstråhle et al. 2020a). Aiming to develop a package that allows the user to visualise multiple experiments in conjunction to create a 3D view of tissue, STUtility builds on well-established methods of analysis (moulded by those established for scRNA-Seq analysis) to focus on novel data visualization (Bergenstråhle et al. 2020a). Highlighting the importance of data normalisation and transformation to deconvolute technical noise from meaningful biological insight, the package uses a regularized negative binomial regression model successfully implemented in Seurat for normalisation (Bergenstråhle et al. 2020a). The image processing capabilities of STUtility focus on the alignment, automatic or manual, of multiple samples in addition to the removal of background noise (Bergenstråhle et al. 2020a). The removal of background noise – called masking in the study – is an integral part of image processing and allows the inside and outside of the tissue to be defined as well as decreasing the images' storage requirements (Bergenstråhle et al. 2020a). To automatically align multiple samples, the package identifies a reference image, then uses an iterative closest point (ICP) algorithm to align the remaining samples to the reference, which can then be reconstructed into a 3D tissue model (Bergenstråhle et al. 2020a). While this method of creating a 3D model is not one which yields the most precise cell segmentation, this trade-off yields greater computational efficiency and still gives a faithful reconstruction of tissue morphology (Bergenstråhle et al. 2020a). Implementation of k-means clustering algorithms allows the package to clearly define the boundaries of the tissue (Bergenstråhle et al. 2020a). For the sequencing data, STUtility leans heavily on the functions created by the package Seurat (Bergenstråhle et al. 2020a). A decomposition of the normalised gene data called non-negative matrix factorization (NMF) is used to choose gene drivers and create a low dimensional representation of the data to be used in defining clusters and nearest neighbours (Bergenstråhle et al. 2020a). To obtain genes whose expression demonstrates spatial patterns, a connection network is created for each spot which allows the package to calculate the spatial-lag of each gene across spots. This is one of the inputs – the other being the normalised counts – used to calculate spatial correlation across the sample (Bergenstråhle et al. 2020a). Its ability to visualise spatial distinct features is clearly demonstrated in determining the spatial relation of gene expression to tissue areas (e.g., a tumour). STUtility is also able to identify SVGs using neighbourhood networks, but its accuracy in performing this function is not compared to other packages (Bergenstråhle et al. 2020a). Other capabilities were tested on a variety of human and mouse tissues (Bergenstråhle et al. 2020a). For both mouse brain and human breast cancer tissue samples, spatial gene expression patterns can be clearly identified (Bergenstråhle et al. 2020a). STUtility allows for the manual alignment of multiple images; however, a comparison as to the accuracy of this method compared to the automatic alignment is not offered and depending on the expertise of the user may vary significantly (Bergenstråhle et al. 2020a). Furthermore, while its implementation of neighbourhood networks offers a promising method to define subsections within a tissue and the heterogeneity within, as would be beneficial during the study of tumours, to see how well this correlates to the heterogeneity of the actual tissues of the sample is not reported (Bergenstråhle et al. 2020a; Palla et al. 2021).

7.2.3 *Assigning Lost Transcripts*

Other packages have been developed with the aim of addressing gaps in analysis that have not been adequately accounted for; one such package is Sparcle (Prabhakaran et al. 2021). When attempting to obtain an accurate gene counts matrix from image-based spatial transcriptomics techniques, often many transcripts are not assigned to cells after segmentation is performed, leading to a loss of data (Prabhakaran et al. 2021). Sparcle aims to recapture the data from these ‘dangling’ transcripts (Prabhakaran et al. 2021). Developed to be used in conjunction with data from any smFISH technology, Sparcle can build a probabilistic model which allows assignment of these dangling transcripts to the appropriate neighbouring cells using a maximum likelihood estimation (MLE). The MLE considers the dangling mRNA’s distance to other transcripts, nearby cells and genes' covariance when calculating which nearby cell the transcript should most accurately be assigned to (Prabhakaran et al. 2021). Similar to other packages, Sparcle assumes that the most accurate representation of gene expression can be modelled using a multivariate Gaussian distribution (Svensson et al. 2018; Prabhakaran et al. 2021). Sparcle can employ two clustering methods when it first groups the cells in the chosen field of vision (FOV) by cell type based on a global count matrix: DPMM and Phenograph. Phenograph is an algorithm developed to cluster cell phenotypes in high-dimensions single-cell data and was originally applied to data from acute myeloid leukemia (Levine et al. 2015). Dirichlet process mixture model (DPMM) is a stochastic process which can feature all the individual Gaussian distributions for the expression of each gene and allows Sparcle to model all these distributions (Neal 2000). While having the additional flexibility to employ either algorithm at the clustering step, during its validation, Sparcle reports data based on the Phenograph algorithm but not on the performance when using DPMM, nor does it specify in which instance one method should be favoured over another (Prabhakaran et al. 2021). When used to assign dangling transcripts to a MERFISH data set, Sparcle was able to assign 68% or almost 2 million missed transcripts, and validation with scRNA-Seq data confirmed that the proportion of cell types assigned post use of Sparcle more closely matched the scRNA-Seq data (Prabhakaran et al. 2021). Validation against other neuronal data sets returned similarly desirable results. Despite this, there are limitations to the use of Sparcle. For example, when the programme draws an area around each dangling transcript that should mimic the size of a cell, the size of this area is optimised to the size of an average neuronal cell, meaning the package might not be well suited to non-neuronal data (Prabhakaran et al. 2021). Sparcle can run on approximately 80 cells in under 10 min with impressive mRNA recovery over three iterations; however, additional data on how this would scale with larger data sets is lacking, potentially causing computational bottlenecks in bigger data sets (Prabhakaran et al. 2021). It claims to improve on packages that remove the cell segmentation step entirely, such as Baysor and SSAM, by removing the need for *a priori* knowledge of the data set and not assuming that the cellular mRNA can be modelled by a uniform distribution (Prabhakaran et al. 2021). However, some further improvements could be made to enhance the performance, such as staining cellular membranes to better understand the size of neighbouring cells rather than estimating based on an area around the nucleus and calculating an estimate of the prior distribution of a gene's localised transcripts (Prabhakaran et al. 2021).

7.2.4 Estimation of Cell Type Composition

Identifying SVGs was the primary focus of the initial packages developed, but it is important to note that packages with alternative aims are increasingly being published. For example, SpatialDWLS was created to improve the identification of different cell types at locations in the data sets which do not have single-cell resolution

(Dong and Yuan 2021). This is termed cell type deconvolution (Dong and Yuan 2021). Other published packages have been developed for this aim, but SpatialDWLS claims to improve on the results of these packages (Dong and Yuan 2021). How SpatialDWLS performs cell type deconvolution can be summarised in two steps: the first uses a cell type enrichment analysis method to identify which kinds of cells have a high probability of being at each location, and the second uses an extension of the dampened weighted least squares (DWLS) method to pinpoint the precise composition of cell types at the specified location (Dong and Yuan 2021). Firstly, signature genes can either be supplied by the user to be identified by differential expression analysis (Dong and Yuan 2021). Building on the previously developed DWLS method for scRNA-Seq data, this was extended to SRT data by incorporating the signature genes step (Dong and Yuan 2021). Furthermore, SpatialDWLS builds on clustering and gene marker identification used in Giotto (Dong and Yuan 2021; Dries et al. 2019). This would imply that any shortcoming with Giotto's performance in these areas would be transferred to SpatialDWLS. When evaluated on a simulated spatial transcriptomics dataset, SpatialDWLS outperformed RCTD and stereoscope in terms of having a lower Root Mean Square Error (RMSE) and in terms of computational time (Dong and Yuan 2021). However, when its performance was tested against a real mouse brain Visium data set, SpatialDWLS's performance was not benchmarked against the other three packages, thus making its performance on real data unclear (Dong and Yuan 2021). Despite this, the authors reported that the spatial location of the cell types assigned by SpatialDWLS was consistent with those reported in the Allen Mouse Brain Atlas (Dong and Yuan 2021). An interesting application of this package was to identify the change of cell type organisation in a spatial-temporal context throughout embryonic heart development (Dong and Yuan 2021). In addition to quantifying an increase in ventricular cardiomyocytes and smooth muscle cells as time went on, by calculating the assortativity coefficient (here used as a measure of whether neighbouring cells were of the same type) the study was able to determine that spatial organisation of the developing heart becomes increasingly defined in terms of neighbourhoods of cell types during development (Dong and Yuan 2021).

Assigning cell types to a spatial transcriptomics dataset can be approached more than one way. By incorporating *a priori* knowledge to a probabilistic likelihood function, FICT (FISH Iterative Cell Type assignment) can blend expression and spatial information to assign cell type to spatial transcriptomics data sets (Teng et al. 2021). This is achieved by creating a generative mixture model using a reduced dimensions representation of expression levels through a denoising autoencoder and assigning each cell as cell type defined by its neighbourhood (represented in an undirected graph); the parameters of this model can be learnt by an expectation maximization approach, which is an iterative process (Teng et al. 2021). Finally, the cell can be classified by a posterior distribution of the model (Teng et al. 2021). During this process, the problem of over-reliance on expression data needs to be addressed, which occurs because in a dataset it is likely that there are more genes being expressed than cell types present (Teng et al. 2021). To circumvent this problem, a named power factor acts as a weight term to balance the dimensionally reduced expression component with the spatial component (Teng et al. 2021). The package was validated using three simulated and real data sets and compared to the results of GMM, scanpy, Seurat and smfishHmrf (Teng et al. 2021). Across all three simulated data sets, FICT has the highest median accuracy, reaching a high of approximately 0.89 in one of the simulated data sets (Teng et al. 2021). When evaluated on a real MERFISH mouse hypothalamus data set, the ground truth of the location of different cell types is unavailable, so clustering results obtained from different animals are compared using the Adjusted Rand Index. When comparing across this metric, FICT is more consistent in applying clusters to the majority of the paired animals, indicating its superior performance in assigning cell type clusters (Teng et al. 2021). FICT has the potential to identify novel subclusters within the data set (Teng et al. 2021). However, FICT's performance drops when applied to data sets with smaller numbers of cells, although this is observed across all packages validated (Teng et al. 2021). Furthermore, its decreased performance was still in line with packages with similar functions, and as spatial transcriptomics data sets become larger, this should not interfere with FICT being applied in future (Moncada et al. 2020). However, despite its greater accuracy when applied to larger datasets, FICT's runtime in these instances could still be improved (Moncada et al. 2020).

RCTD is another package created with the final aim of identifying cell types in a spatial transcriptomics data set (Cable et al. 2020). While identifying SVGs is extremely informative, it is important to understand how the role of underlying cell types contributes to a gene's spatially variable expression patterns (Cable et al. 2020). Robust Cell Type Decomposition (RCTD) makes use of annotated scRNA-Seq data to create cell type profiles for expected cell populations in the data, then labels spatial transcriptomics pixels with cell types using a supervised learning method (Cable et al. 2020). As one of the major hurdles in this analysis is the fact that the current spatial transcriptomics data sets can contain multiple cell types within a single pixel, RCTD can also fit a statistical model to determine multiple cell types present within a pixel and normalise across platform effects between the scRNA-Seq and SRT datasets (Cable et al. 2020). To achieve this, RCTD first creates a spatial map of cell types and estimates the number of different cell types in each pixel where the gene counts are assumed to have a Poisson distribution (Cable et al. 2020). This should circumvent the problem introduced by the current unsupervised learning methods that overlook clustering cells that co-localise transcriptionally as well as spatially (Cable et al. 2020). Using this approach, RCTD was able to classify cells across platforms with almost 90% accuracy. However, as with any supervised learning approach, the cell types one can detect using this tool are limited to how accurately and fully the reference data set is annotated, which may present difficulties. Also, while the study tested RCTD using references and data sets generated by many different kinds of scRNA-Seq and SRT technology, the effects that specific platforms may have on cell type assignment is still undetermined.

7.2.5 Spot-by-Spot Clustering

A common step in the analysis of many kinds of omics data sets is to perform clustering, and this is prevalent when analysing SRT data. This section will discuss techniques that cluster spots on an SRT array, which may contain multiple cell types, based on the overall gene expression profile of the spot (Bergenstråhle et al. 2020b). Despite being common, this is not a straightforward step. Understanding the results after different iterations can prove difficult, as does choosing the correct hyperparameters (Bergenstråhle et al. 2020b). This is further confounded as each barcode is associated with multiple cells (Bergenstråhle et al. 2020b). To address these issues, an R package called SpatialCPie was developed which focuses on clustering spots on the array based on the gene expression profile to allow annotation of regions of the tissue (Bergenstråhle et al. 2020b). SpatialCPie allows the user to choose which algorithm to implement and clusters the data at different resolutions from the start (Bergenstråhle et al. 2020b). The user is then free to choose which conformations of clusters created at which resolution most accurately represent their data. By creating a cluster graph and an array plot, SpatialCPie gives the user varied insight into how different resolutions affect the clustering outcomes (Bergenstråhle et al. 2020b). The cluster graph displays how the different clusters relate to one another across different resolutions, and conveys the origins of new clusters as they emerge at higher resolutions (Bergenstråhle et al. 2020b). The edges of the graph link the percentage of spots in new clusters that descend from different lower resolution clusters (Bergenstråhle et al. 2020b). The second visualisation method is the array plot, which represents the SRT array, but each spot is depicted as a pie cart that shows how similar the gene expression is between cluster centroids and spatial regions (Bergenstråhle et al. 2020b). SpatialCPie offers the novel, to the best of the authors' knowledge, option to choose a particular region of the dataset for further sub-clustering which may be appropriate depending on the tissue of interest (Bergenstråhle et al. 2020b). While SpatialCPie only compares itself to ST viewer – in a limited capacity – its overall performance is promising (Bergenstråhle et al. 2020b). However, additional validation of its performance compared to other similar packages such as ST viewer would be beneficial to understand its accuracy.

7.2.6 Pipelines

As the area of SRT continues to expand, pipelines, rather than just analysis packages, will become more commonplace. One of the first available pipelines written in R is Giotto, which is a platform that can be used on both transcriptomics and proteomics data; it is divided into a data analysis and visualisation module (Dries et al. 2019). With a focus on being user-friendly and reproducible, Giotto does provide the opportunity for more complex spatial analysis using HMRF models (Dries et al. 2019). As a foundation, Giotto creates a neighbourhood network of cells and a spatial grid for downstream analysis which includes ligand-receptor identification, gene expression pattern analysis and determining preferential cell neighbours (Dries et al. 2019). Giotto is tested on ten different data sets obtained with varying technologies and from varied tissues to examine its performance across a range of benchmarks (Dries et al. 2019). The initial steps in the analysis are similar to those performed in scRNA-Seq analysis, but Giotto does offer three different algorithms for identifying marker genes, one of which (Gini) was specifically developed for the pipeline, which differ in their strength in identifying particular kinds of marker genes (Dries et al. 2019). The Scran method evaluates the markers between two groups of cells by running t-test (default) and then determining marker genes (Lun et al. 2016). Mast identifies marker genes between two cell groups by employing a hurdle model (Finak et al. 2015). The Gini algorithms score marker genes within a cluster based on Gini coefficients, which were developed to identify rare cell types from an adapted model implemented in the social sciences (Jiang et al. 2016). All of these algorithms were developed to score marker genes between clusters in single-cell data sets. When evaluated, Gini discovered the most marker genes for the 12 cell types when compared to Mast and Scran; however, when identifying the top 20 markers using each method, Gini had the lowest sensitivity but highest specificity in both the endothelial and oligodendrocyte populations (Dries et al. 2019). The sensitivity and specificity of each algorithm vary slightly across the different cell populations they investigated when evaluated against a sequential fluorescence in situ hybridization (seqFISH+) somatosensory cortex dataset, and this is important when deciding which algorithm to employ; furthermore, this needs to be tested against data sets generated from different biological material and technologies to best understand the true limitations of each algorithm (Dries et al. 2019). Giotto also has analysis pipelines designed specifically for SRT data sets with lower resolution (Dries et al. 2019). By using one of three algorithms to provide an enrichment score between a location's expression pattern and a cell's gene signature, it is possible to assign a cell type to a location which contains more than one cell (Dries et al. 2019). Once again, the availability of multiple algorithms at this step which require different inputs allows Giotto to be flexibly implemented on a number of different datasets (Dries et al. 2019). These three enrichment algorithms were validated on a simulated dataset similar to one generated using seqFISH+ with the hypergeometric algorithms having the lowest AUC score (0.8) and both PAGE and RANK scoring similarly well when predicting cell type at a particular location (Dries et al. 2019). When applied to real data sets, the two best scoring algorithms RANK and PAGE performed well and should be used when employing the Giotto pipeline (Dries et al. 2019). To analyse spatial patterns of gene expression, Giotto creates a spatial network to represent the data using a Delaunay triangulation network, which is the same as the method employed by scGCO (Zhang et al. 2018; Dries et al. 2019). While the option is available to alternatively construct a spatial network with two different methods offering the user greater control on downstream parameters, the analysis results appear insensitive to these adjustments (Dries et al. 2019). To uncover SVGs, Giotto introduces two new methods, BinSpect-kmeans and BinSpect rank, as well as incorporated methods from SpatialDE, Trendsceek and SPARK

(Dries et al. 2019). When evaluated, each of the methods identified unique SVGs, with 103 genes being identified by all five methods (Dries et al. 2019).

As the field of SRT continues to expand, so will the analytical tools available. As an increasing number of downstream analysis packages are published for SVG identification amongst other analyses, pipelines and frameworks will become increasingly complex in the scope of their abilities. A new framework developed to combine and encompass all aspects of analysis for spatial-omics technology is Squidpy (Palla et al. 2021). While not built specifically for the analysis of SRT data, the Squidpy framework developed in Python brings common tools for analysis and visualisation to any spatial-omics data and takes advantage of the additional information available to improve exploration (Palla et al. 2021). Offering a broader and more modular approach than Giotto, Squidpy offers the opportunity for other packages to be easily integrated into its pre-existing framework to expand its capabilities (Palla et al. 2021). Squidpy will store the image data in an Image Container and create a neighbourhood graph of spatial coordinates so that it can be used on a wide array of technologies (Palla et al. 2021). A feature of Squidpy that adds additional analytical opportunity is its in-built image analysis tools (Palla et al. 2021). While the packages discussed so far require an image as part of the input for analysis, none extend so far as to allow the user to investigate the data contained in this image to the same extent as Squidpy, which is the capability that differentiates it most from Giotto (Palla et al. 2021). The first step in the investigation of cellular neighbourhoods and spatial patterns is the construction of a spatial graph (Palla et al. 2021). When compared to similar processes in Giotto, Squidpy had a more efficient run time when constructing both a spatial graph and calculating neighbourhood enrichment, although for data sets with a smaller number of observations the difference was not great (Palla et al. 2021). Despite offering an interesting perspective on the direction of spatial-omics analysis frameworks and pipeline and reporting limited but promising results with regards to its ability to reproduce results about cellular neighbourhoods, Squidpy does not report its performance in accurately discovering SVGs nor does it quantify how its results relate to those reported in the previous studies (Palla et al. 2021).

7.2.7 Discussion

Despite being a relatively novel technology, SRT – often alongside scRNA-Seq or other techniques – has already been successfully applied to identify gene expression changes in a variety of tissues and disease states. One example was its application in mouse brains to understand spatially DEGs involved in early-stage Alzheimer’s disease (Navarro et al. 2020). Different SRT methods are best suited to studying different cell types within a tissue to distinguish differences between them in disease states, such as comparing the dopamine neurons from two regions in Parkinson’s patients (Aguila et al. 2018). To further demonstrate how this technology can be applied to an array of conditions and diseases, Modlin and colleagues successfully actioned it as part of an investigation into the organisation of cellular subtypes that contribute to the antimicrobial capabilities of human leprosy granulomas (Ma et al. 2020).

This clear increase in the popularity of SRT has prompted the recent development of many different packages and pipelines for the downstream data analysis of SRT data sets. While it seems that certain studies are still reliant on packages developed for scRNA-Seq data adapted to included SRT analysis such as Seurat (Ortiz et al. 2019), the variety of purpose-built available tools will likely replace these. A package for easily identifying SVGs seems to be the most popular aim, and even the pipelines developed so far have centred around this same purpose (Svensson et al. 2018; Li et al. 2021; Sun et al. 2019; Edsgärd et al. 2018; Hao et al. 2021; Zhang et al. 2018; Palla et al. 2021; Dries et al. 2019). However, the scope of developing packages continues to expand to further improve the capabilities of analysis, such as Sparcle, which was developed to be used in conjunction with other packages.

Of all the packages discussed, SpatialDE seems to be the most popular, followed by SPARK, Trendsceek and Giotto in terms of being used as benchmarks by which to validate new packages. SpatialDE indicated a tendency to label genes with very low expression as SVGs (Sun et al. 2019), and certain discrepancies in performance compared to other packages tested on real data sets. This alongside the potential introduction of false positives indicates an area of improvement for this popular package. A current limitation of the validation of package performance is that most commonly two data sets (Ståhl et al. 2016), obtained using the same Visium method, are used which will surely introduce inherent bias to the benchmarking process. It would be beneficial to understand the package's performance across datasets from different tissues (instead of exclusively olfactory bulb and breast) generated using a different technology.

To most comprehensively establish the relative performance of all packages, a review should be conducted which benchmarks all packages simultaneously against the same datasets, generated by different SRT methods in different tissues and a standard method for validation established. More packages that are modular and can be integrated alongside one another to expand the scope of analysis are critical and will help advance the field and uptake of this technology. Additionally, the further development of user-friendly pipelines will also make analysing SRT results more accessible. As the array of available tools for analysis of SRT data becomes greater, the results from studies employing the technology will improve and the scope of biological problems that can be addressed will simultaneously expand.

Declarations **Consent for Publication**: All authors provide their consent for publication.

Competing Interest: The authors declare no competing financial interest.
Authors' Contributions: NC drafted the manuscript with intellectual input contributions from HTN and MR. HTN and MR reviewed the manuscript. All authors approved the final manuscript.

References

Aguila J et al (2018) Spatial transcriptomics identifies novel markers of vulnerable and resistant midbrain dopamine neurons. bioRxiv. https://doi.org/10.1101/334417

Asp M, Bergenstråhle J, Lundeberg J (2020) Spatially resolved transcriptomes—next generation tools for tissue exploration. BioEssays 42:1900221

Bergenstråhle J, Larsson L, Lundeberg J (2020a) Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics 21:482

Bergenstråhle J, Bergenstråhle L, Lundeberg J (2020b) SpatialCPie: an R/Bioconductor package for spatial transcriptomics cluster evaluation. BMC Bioinformatics 21:161

BinTayyash N et al (2020) Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. bioRxiv 2020.07.29.227207. https://doi.org/10.1101/2020.07.29.227207

Cable DM et al (2020) Robust decomposition of cell type mixtures in spatial transcriptomics. bioRxiv 2020.05.07.082750. https://doi.org/10.1101/2020.05.07.082750

Chen G, Ning B, Shi T (2019) Single-cell RNA-seq technologies and related computational data analysis. Front Genet 10:317

Crosetto N, Bienko M, Van Oudenaarden A (2015) Spatially resolved transcriptomics and beyond. Nat Rev Genet 16:57–66

Dong R, Yuan G-C (2021) SpatialDWLS: accurate deconvolution of spatial transcriptomic data. bioRxiv 2021.02.02.429429. https://doi.org/10.1101/2021.02.02.429429

Dong J et al (2018) Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. Genome Biol 19:31

Dries R et al (2019) Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. bioRxiv. https://doi.org/10.1101/701680

Edsgärd D, Johnsson P, Sandberg R (2018) Identification of spatial expression trends in single-cell gene expression data. Nat Methods 15:339–342

Exelby K et al (2021) Precision of tissue patterning is controlled by dynamical properties of gene regulatory networks. Development 148:dev.197566

Finak G et al (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 16:278

Hao M, Hua K, Zhang X (2021) SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. bioRxiv. https://doi.org/10.1101/2020.12.10.419549

Harzing AW (2016) Publish or perish? Harzing.com https://harzing.com/resources/publish-or-perish/os-x. Accessed 26 Apr 2021

He S et al (2020) Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. Genome Biol 21:294

Hu J et al (2020) Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. bioRxiv 2020.11.30.405118. https://doi.org/10.21203/RS.3.RS-119776/V1

Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50:96

Jiang L, Chen H, Pinello L, Yuan GC (2016) GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol 17:144

Karaayvaz M et al (2018) Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun 9:1–10

Lein E, Borm LE, Linnarsson S (2017) The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. Science 358:64–69

Levine JH et al (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell 162:184–197

Li Q, Zhang M, Xie Y, Xiao G (2021) Bayesian modeling of spatial molecular profiling data via Gaussian process, Bioinformatics, 37(22):4129–4136, https://doi.org/10.1093/bioinformatics/btab455

Lun ATL, McCarthy DJ, Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Research 5:1–71

Ma F et al (2020) Single cell and spatial transcriptomics defines the cellular architecture of the antimicrobial response network in human leprosy granulomas. bioRxiv 12.01.406819. https://doi.org/10.1101/2020.12.01.406819

Marx V (2021) Method of the year: spatially resolved transcriptomics. Nat Methods 18:9–14

Moncada R et al (2020) Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol 38:333–342

Navarro JF et al (2020) Spatial transcriptomics reveals genes associated with dysregulated mitochondrial functions and stress signaling in Alzheimer disease. iScience 23:1–19

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9. http://www.jstor.org/about/terms.html

Ortiz C et al (2019) Molecular atlas of the adult mouse brain. bioRxiv. https://doi.org/10.1101/784181

Palla G et al (2021) Squidpy: a scalable framework for spatial single cell analysis. bioRxiv 2021.02.19.431994. https://doi.org/10.1101/2021.02.19.431994

Prabhakaran S, Nawy T, Pe'er' D (2021) Sparcle: assigning transcripts to cells in multiplexed images. bioRxiv 2021.02.13.431099. https://doi.org/10.1101/2021.02.13.431099

Regev A et al (2017) The human cell atlas. elife 6:e27041

Shah S, Lubeck E, Zhou W, Cai L (2016) In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron 92:342–357

Ståhl PL et al (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353:78–82

Sun S, Zhu J, Zhou X (2019) Statistical analysis of spatial expression pattern for spatially resolved transcriptomic studies. bioRxiv. https://doi.org/10.1101/810903

Svensson V, Teichmann SA, Stegle O (2018) SpatialDE: identification of spatially variable genes. Nat Methods 15:343–346

Teng H, Yuan Y, Bar-Joseph Z (2021) Cell type assignments for spatial transcriptomics data. bioRxiv 2021.02.25.432887. https://doi.org/10.1101/2021.02.25.432887

Tiklová K et al (2019) Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development. Nat Commun 10:1–12

Waylen LN, Nim HT, Martelotto LG, Ramialison M (2020) From whole-mount to single-cell spatial assessment of gene expression in 3D. Commun Biol 3:1–11

Ximerakis M et al (2019) Single-cell transcriptomic profiling of the aging mouse brain. Nat Neurosci 22:1696–1708

Zhang K, Feng W, Wang P (2018) Identification of spatially variable genes with graph cuts. bioRxiv 491472. https://doi.org/10.1101/491472

Zhou S et al (2021) Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks. Mol Ther Nucleic Acids 23:682–690

Chapter 8
The Interplay Between the Transcriptomics and Proteomics Profiles

John Oluwafemi Teibo, Virgínia Campos Silvestrini, Alessandra P. Vargas, Guilherme Pauperio Lanfredi, and Vítor Marcel Faça

8.1 Introduction

Living organisms have complex physiology, with extremely regulated systems to modulate responses to internal and external stimuli, allowing adaptability in the environment in which they live. These processes involve constant synthesis and degradation of biomolecules as a response to cellular events. The same occurs in pathological situations, where the abnormal stages of development of a disease are carried out by important changes in the set of biomolecules responsible for cell or organism function. Genetic products, mainly mRNAs and proteins, are constantly being modulated in response to normal physiological and pathological cellular events. Therefore, effective monitoring of the cellular or organism complement of mRNAs and proteins, namely the transcriptome and proteome, respectively, are fundamental to understand normal as well as pathological molecular mechanisms in the cells.

Technological advances in genomic sciences, including modern tools for transcriptomics and proteomics, have been recognized as important drivers of biosciences, allowing significant scientific discoveries and biological advances in the last decade. The genomics, transcriptomics, and proteomics now can routinely provide information on cell mutations, disease biomarker, gene therapies, with particular direct impact in personalized medicine applications. More importantly, these approaches can be used coordinately in the same study, providing deep molecular profiles in healthy and diseased situations (Manzoni et al. 2018).

Great scientific advances started with genomics. However, despite being revolutionary in the beginning of the century, the study of the genome does not respond to

J. O. Teibo · V. C. Silvestrini · A. P. Vargas · G. P. Lanfredi · V. M. Faça (✉)
Department of Biochemistry and Immunology and Center for Cell Based Therapy – Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil
e-mail: vitor.faca@fmrp.usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_8

all the challenges and questions posed by biology especially in human health and disease areas. Many regulatory mechanisms are orchestrated between the genome transcription and the translation of proteins, which are responsible for most of the phenotypic characteristics of an organism (Buccitelli and Selbach 2020). Based on this principle, other technologies and approaches emerged with the aim of identifying a complete set of transcripts and also proteins in a given biological system. The integration of these "omics" approaches can be the key to understanding complex data, helping to generate more complete hypotheses in several areas of biology.

From the earlier studies comparing mRNA and protein abundancy in some biological models, it was obvious the lack of full concordance observed in high-throughput experiments (Anderson and Seilhamer 1997; Gygi et al. 1999). On the biological front, differences could be initially attributed to RNA splicing, differential RNA and protein turnover, post-translational modifications, allosteric protein interactions, and proteolytic processing events. On the experimental front, challenges in experimental design and data interpretation, as well as technological limitations, contributed to some of the differences observed (Hegde et al. 2003).

In this chapter, we will explore the biological and technical factors that affect the concordances and differences already established for the interplay of transcriptomes and proteomes in biological system. After a couple of decades with scientific and technological advances in the "omics" field, researchers have been elucidating some new players and factors responsible for the imbalance between mRNA and proteins, some of the technical limitations has been overcome, data generation and processing and more importantly, how and when data from both transcriptomics and proteomics can be integrated have been improved and properly compared in order to draw hypothesis and conclusions that now can form the basis of pathways involved in health and diseases processes in various biological systems.

8.2 Transcriptomics

The analysis of the entire transcriptome (mRNA, tRNA, rRNA, and miRNA) has become an essential tool in the quantification of gene expression in different tissues, organs, and cells, previously identified only by DNA sequencing. Transcriptomics can routinely provide an overview of the characteristics of gene expression in different samples, determining the presence/absence and quantification of transcripts. These transcripts profiles provide the basis for understanding regulatory pathways that control cell function, growth, and development in different biological systems. This information is also essential for understanding the metabolic and tissue dynamics, especially in comparisons among physiological and pathological states (Jiang et al. 2015).

The dominant contemporary techniques used for transcripts profiling are RNA microarray and RNAseq, both with their distinct advantages and disadvantages. Microarrays measure the abundances of an established set of transcripts via their hybridization with an array of complementary probes, allowing the analysis of thousands of transcripts at a low cost (Lowe et al. 2017). The abundance of transcription is determined by hybridization of fluorescently labeled transcripts to these know probes. Importantly, this approach is based on a defined set of known sequences, to generate the profiles for the array (Barbulovic-Nad et al. 2006; Lowe et al. 2017). Current commercially available mRNA microarrays can profile virtually the entire transcriptomes for several different organisms.

More recently, RNAseq, which refers to the complete sequencing of the transcriptome, determines the abundance of mRNA from the number of counts from each transcript (Morozova et al. 2009). The first paper published using this technique was in 2006 with 105 transcripts sequenced and that provided sufficient sequence coverage to quantify their relative abundance (Bainbridge et al. 2006). RNAseq became more established and robust from 2008 with the emergence of next generation sequencing and massive sequencing by synthesis (SBS) technology, which now is sufficient for accurate quantitation of the entire human transcriptome (Lappalainen et al. 2013). These approaches have led to rapid expansion of this technology to answer many biological questions revolving around the transcriptomes for health and diseased biological problems.

For both the microarray and RNAseq strategies, it is initially necessary to purify the RNA and convert it into complementary DNA (cDNA). Subsequently, cDNAs are chemically marked with fluorophores and hybridized to probes on the chip to detect present target genes, if the technique used is RNA-microarray. On the other hand, if the strategy used is that of RNAseq, the RNA can also be fragmented to build a library for sequencing analysis. Both strategies must be executed through the platform of choice according to the specific objectives of each experiment or study (Nagalakshmi et al. 2010; Kukurba and Montgomery 2015; Manzoni et al. 2018).

Microarrays are now a robust technique and they are commercially available for complete genomic coverage using optimized sets of probes. However, transcripts not included in the probes will not be observed. More complex organisms have a greater number of exons and also non-coding sequences (introns). In this case, direct sequencing of mRNA molecules can provide more information about transcription products and potential translation products with greater coverage. Based on this principle, RNAseq does not require prior knowledge of the transcripts in the sample, making it possible to compare different sets of genes. This unique feature has enhanced the discovery of novel transcript products from the gene expression process. The microarray technique, on the other hand, has a lower cost and allows a higher number of replicates, necessary for confident new discoveries (Nagalakshmi et al. 2010).

The integration of transcriptomics with other "omics" technologies can help to grasp the complexity of cell life. Transcriptomics tools permit the parallel quantification of thousands of biomolecules and therefore allow for explorative, non-hypothesis-driven studies. However, there are several sources of variability originating from biological and technical causes that can affect the quality of the resulting data, such as biological heterogeneity in the sample, sample collection variations, RNA quantity and quality obtained from preparation steps, technical variation during sample processing, and batch effects, among others. Some of these issues can be avoided with an appropriate and carefully experimental design that controls for the different sources of variation, but others will be detected only after a quality assessment of the raw data through computational support tools. Therefore, regardless of the technology used to measure gene expression of a cell, ensuring quality control is a critical starting point for any subsequent analysis of the data (Cobb et al. 2005; Larkin et al. 2005; Irizarry et al. 2005; Heber and Sick 2006). As we will discuss ahead, other technologies such as proteomics also present advantages and technical limitations. These potential limitations are the initial key factors to be eliminated or at a least minimized to allow multi-omics platforms data integration and comparison to provide more information toward answering complex biological questions.

8.3 Proteomics

Proteomics consists of suite of techniques that allow proteome analysis, making possible to identify proteins and quantify their abundance and post-translational modifications (PTMs) in a complete and complex set of different samples, including cells, tissues, and fluids, among others (Faça 2017). Unlike the transcriptomics strategies mentioned above, proteomics provides direct measurements on active and post-translationally modified proteins, in addition to their cell expression and localization. This kind of information is essential during the development of several pathologies, since biochemical processes such as splicing, phosphorylation, ubiquitination and other PTMs are usually severely impacted. Thus, proteomics studies also provide information on altered pathways, contributing to the discovery of important biological targets in the emergence of diseases (Silvestrini et al. 2019). Therefore, proteomic strategies are important to complement genomic and transcriptomic information (Aslam et al. 2017; Silvestrini et al. 2019).

When studying the proteome, an increase in the degree of molecular complexity in relation to the genomic study must be considered. The four-nucleotide codes of DNA and mRNA are translated into a complex code of 20 amino acids with different combinations, forming primary sequences that can adopt specific chemical conformations and modifications to produce a functional protein (Manzoni et al. 2018). The proteome is a multidimensional and highly dynamic system, in which each protein has several interconnected properties that together represent the phenotype of a cell or organism.

Although some of the underlying technology for quantifying protein abundance was introduced more than 40 years ago (O'Farrell 1975; Klose 1975), there has been recently a significant advance in the field and the development of new tools. With the advances in mass spectrometry which focus on studying proteomes, cell location, synthesis/degradation, Post-translational modifications, (PTMs) etc. has began to be analyzed in an integrated manner, allowing a better understanding of physiological and cellular processes (Larance and Lamond 2015).

The need to understand cellular changes at the protein level has led to the emergence of more accurate, high-quality proteomic strategies that guarantee sensitivity for the simultaneous identification and quantification of thousands of proteins in a sample. Most common proteomic studies in disease development are based on liquid-chromatography coupled with high-throughput mass spectrometry (LC-MS/MS) technology. In particular, LC-MS/MS has enabled the structural characterization of proteins and protein complexes that have been intractable through other methods, providing experimental evidence with high resolution (Chandramouli and Qian 2009). Currently, two principles of analysis are used: a global and targeted proteomics approach (Chandramouli and Qian 2009). In global proteomics or also called shotgun sequencing or whole proteome analysis, there is no hypothesis of specific proteins to be found in the sample. These approaches have gained interest in clinical applications since a high number of altered proteins are observed in different conditions, and they can be evaluated in a quantitative manner, using a wide range of approaches that are mainly divided in isotopic-labeling or label-free methods (Chandramouli and Qian 2009; Silvestrini et al. 2019, 2020). This strategy currently provides a detailed map of thousands of proteins and their respective abundance, allowing comparison of few different variables in each experiment.

Conversely, in targeted proteomics, a known and specific set of proteins are quantitatively analyzed by mass spectrometry. Panels are created with unique peptide sequences that represent the target protein and that will be accurately monitored during the experiment. This approach is based on the high selectivity of peptide ions filtering to improve the sensitivity and accurate quantification of ions (Faça 2017; Silvestrini et al. 2019). Also, this approach allows faster methods and larger number of samples and variables analyzed, which is still a limitation for high-throughput/shotgun proteomics. In summary, the combination of shotgun and targeted strategies provides additional capabilities to identify and validate protein molecular signatures, for example in patient sample cohorts, since the global analysis identifies the altered proteins and subsequent individual sample are accurately quantified for the set of selected proteins by targeted proteomics (Lanfredi et al. 2021). This is a potent strategy for the discovery of new pathways and disease molecular signatures in various perturbed conditions for a wide variety of biological systems.

8.4 Mechanisms That Regulate mRNA and Protein Levels

For many years, the central dogma of molecular biology stated that RNAs molecules were intermediates between DNA and protein and that the function of RNA was primarily linked to the translation of the genetic material into polypeptide chains (proteins) (Brenner et al. 1961; Jacob and Monod 1961). Therefore, the basic level of understanding of the central dogma of biology supports that protein concentrations in a biological system should then directly correlate with their respective mRNAs levels, since translation is required to produce proteins. In fact, it has been shown that when mRNA levels are low, usually proteins are not detected and the ability to detect proteins increases significantly at higher levels of mRNA (Vogel and Marcotte 2012).

Considering the last decades' technological advances, the measurement of transcribed mRNA has proven to be very powerful in the discovery of molecular markers and the elucidation of functional biological mechanisms. However, it was also evident that mRNA abundance is not a good predictor of protein abundance in the cell. Many possible points of control and potential interruption for the flow of information coded in DNA sequences until it becomes a functional protein have been elucidated. The synthesis and turnover of cellular proteins require several processes that are interconnected, starting with the transcription, processing, and translation of mRNAs, followed by protein folding, cellular transport and localization, and post-translational modification. Parallel processes such as mRNA degradation, inhibition of mRNA translation and protein degradation also modulate the amount of functional protein available in the cell, tissue or organism, which directly impact their physiological conditions (Vogel and Marcotte 2012). These basic mechanisms are illustrated in Fig. 8.1. Given this high degree of interconnection of processes that affect both mRNA and protein levels, understanding normal physiological

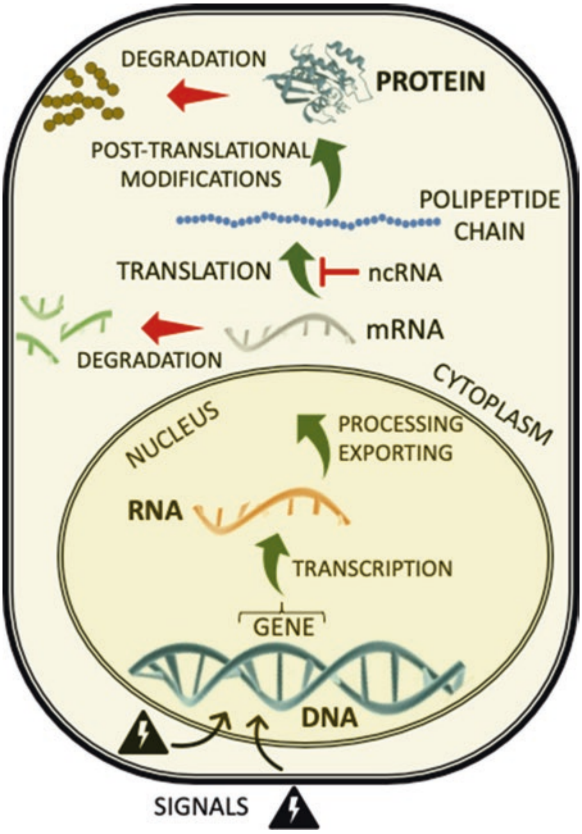


Fig. 8.1 General overview of gene expression. The diagram depicts main processes that are responsible for producing (green arrows) or degrading (red arrows) mRNAs and proteins at a cellular level. Upon intracellular or extracellular signals or stimulus, gene expression is triggered and the fine balance between all these processes is responsible for the maintenance of cellular physiology and defines its normal momentary phenotype

processes that modulate these biomolecules are important to introduce the correlation of mRNA and protein levels in health and disease. Below we discuss some of the most established points in that regard.

8.4.1 mRNA Transcription and Decay

The average abundance of mRNA in a given cell, tissue, or organism is determined by the rates of the transcription versus degradation. Transcriptional regulation occurs at two interconnected levels: the first involves transcription factors and the transcription apparatus, and the second involves chromatin and its regulators. All starts with DNA binding transcription factors that occupy specific sequences and recruit and regulate the transcription by the RNA polymerase II machinery. In eukaryotic systems, there has been extensive study of specific transcription factors and their cofactors, the general transcription apparatus, and various chromatin regulators, leading to current models for specific gene transcription control. The particular set of transcription factors that are expressed in any cell or tissue type at a given moment controls the selective transcription of a subset of genes, correspondent to that cell or tissue expression program. Therefore, the set of genes that are transcribed largely defines the cell phenotype. The gene expression program of a specific cell type includes RNA species from genes that are active in most cells (housekeeping genes) and genes that are active predominantly in one or a limited number of cell types (cell-type-specific genes). Studies of the transcription factors that are key to establishing and maintaining specific cell states suggest that only a small number of the transcription factors that are expressed in cells are necessary to establish cell-type-specific gene expression programs (reviewed by Lee and Young 2013).

On the other end, decay of mRNA can be broadly divided into two classes: mechanisms of quality control that eliminate the production of potentially toxic proteins and mechanisms that lengthen or shorten mRNA half-life for the purpose of changing its abundance, and therefore the availability of functional proteins. Because mRNAs primarily function as templates for protein synthesis, it is logical that cells have evolved translation dependent quality-control mechanisms to dispose of defective mRNAs that synthesize abnormal proteins. Nonsense-mediated mRNA decay (NMD), which, unlike most mRNA decay pathways, appears to be restricted to newly synthesized transcripts, which occurs in all eukaryotes that have been studied, eliminates mRNAs that prematurely terminate translation. This mechanism dampens the potentially toxic effects of defective transcripts that are routinely generated during gene expression of newly synthesized mRNAs. In addition, NMD is inhibited by negative regulators induced by some stresses – such as amino acid starvation and viral infection, among others. Mature mRNAs are degraded by exonucleases acting at both ends of the molecule or endonucleases. Decay rates can be specified by control elements that are usually located within the 3′-untranslated regions (UTRs) of mRNAs and are recognized by various RNA-binding proteins

(RBPs) (Wilusz et al. 2001; Parker and Song 2004). Additionally, degradation of transcripts occurs at distinct cytoplasmic sites in both yeast and human cells indicating that the regulation of mRNA stability is a widespread, tightly regulated, and conserved mechanism for the control of gene expression (Wilusz and Wilusz 2004). Interestingly, little is known about how ribonucleases are regulated, particularly because this class of enzymes is regulated through the proteins they interact with (reviewed by Schoenberg and Maquat 2012).

The time-course measurements of mRNA abundance are, therefore, the key factor to evaluate turnover and stability. Technological advances made the global evaluation of mRNA turnover more common and efficient than it is for proteins. Genome-wide mRNA turnover has been determined in bacteria (Bernstein et al. 2002; Selinger et al. 2003), yeast (Wang et al. 2002; Grigull et al. 2004), plants (Gutiérrez et al. 2002), and humans (Raghavan et al. 2002; Yang et al. 2003) by measuring mRNA levels at different times after RNA polymerase II inactivation. In fact some of these studies brought the concept of timing to describe mRNA stability. Each RNA polymerase II can transcribe about ~100 primary mRNAs per hour from the DNA template. In contrast, ribosomes produce up to 10,000 protein molecules per mRNA per hour (Darzacq et al. 2007; Hausser et al. 2019). We will discuss more about the differences in timing and turnover rates for mRNA and proteins in the next sections.

8.4.2 Regulation of Protein Translation

Cellular functions depend on simultaneous participation of thousands of proteins, which are in a dynamic equilibrium of abundance to maintain homeostasis. As we have been discussing, the cellular processes of protein translation, folding, and degradation together determine the total repertoire of cellular proteins. Protein levels in cells, tissues, and organisms are extremely well regulated in order to reflect the healthy phenotype. Therefore, there should be a very efficient balance between the mechanisms of production and degradation of proteins. In fact, protein translation is the most energy consuming process in the cell, requiring fine modulation before the different stimulus provided by the cellular microenvironment according to the variety of needs of the organism. Starting from the availability of the particular transcriptome of a cell in a given moment, post-transcriptional control takes place during translation, and encompasses both global and transcript-specific mechanisms to regulate protein synthesis (Dever 2002; Gebauer and Hentze 2004). Global regulation, which affects the translation of most transcripts, usually occurs by changes in the phosphorylation state of translation initiation factors and by adjusting the number of available ribosomes (Preiss and Hentze 2003). Transcript-specific regulation, by contrast, modulates the translation of a distinct group of mRNAs and is mediated by a large diversity of mechanisms, such as codon bias or the interaction of the transcript with regulatory elements (Beilharz and Preiss 2004). It involves RNA binding proteins that associate with particular structural features or control elements present in the UTRs of target transcripts, and are similar to the control of RNA decay which is highlighted earlier.

Among the various processes that coordinate the mRNA translation is the mTOR signaling pathway listed as one of the most studied and understood. The mTOR signaling pathway can rely on various external stimuli to continue the translation regulation. Hormones, growth factors, metabolites, and nutrients can start cell translation machinery (Buttgereit and Brand 1995). mTOR is a Ser/Thr kinase that stimulates anabolic processes through the phosphatidylinositol 3-kinase/protein kinase B (PI3K/AKT) signaling pathway activation by hormone or growth factor via specific receptor tyrosine kinase complexes and its specific substrates. Notably, the entire functional control of these pathways is regulated by post-translational modification, mainly phosphorylation. In addition to regulation by external factors, the mTOR translation modulation pathway is also affected by the cell's internal signaling by conditions such as hypoxia and energy depletion. Another way of modulating the translation extensively studied is the MAPK (mitogen-activated protein kinases); this pathway regulates among others translation parallel to mTOR, interacting with it at several points enabling or inhibiting translational activity. Considering the beginning of MAPK pathway, the upstream event begins with Ras GTPases that can be activated by several external stimuli, also interacting with MAPKs that regulate TSC complex to finally affect mTORC1, or downstream with modulation of translation machinery stimulating its components, such as elf4E (Shaw and Cantley 2006). Despite the detailed understanding of signaling pathways for the components responsible for mRNA translation, such as the regulatory role of the PI3K/mTOR and Ras/MAPK pathways, they are not unique, recent efforts with different analytical techniques show the role of additional signaling pathways in the activation of the translational machinery and even of sensitive or specific transcripts of given pathway (Roux and Topisirovic 2018).

With all these roles assigned, the signaling pathways involved in the translation also prove to be a relevant target for the therapy of diseases, since the imbalance in this adjustment has great potential in the appearance of organism disorders. Comprehensive analysis using a proteogenomics approach of the PI3K/AKT/mTOR pathway showed high activity of these pathways in a significant portion of cancers and despite the great correlation of activity rates, there is in some cases decoupling, showing the regulatory character in the multiple levels of these pathways (Zhang et al. 2017).

8.4.3 Non-coding RNAs Inhibit mRNA Translation

We have been discussing many aspects that affect mRNA translation. However, the development and application of deep sequencing have shown that most of the genome results in transcription to RNAs, but from these only 1–2% of the human genome codes for proteins. Hence, it is possible to divide the transcriptome into two large groups, being coding potential RNAs, that have potential to be translated into proteins and RNAs without coding potential, not being translated into proteins, non-coding RNAs (ncRNAs). Even though the RNAs were already studied extensively, currently represented mainly by mRNAs, ncRNAs account for the major part of RNAs, holding a great potential for the knowledge of new mechanisms of the processes of expression (Dunham et al. 2012). In fact, the discovery of microRNAs (miRNAs) in 1993 (Lee et al. 1993) followed by developments and discoveries in small RNA and other ncRNA species have redefined the gene regulation landscape. These RNA molecules play a significant role in modulation of an array of physiological and pathological processes that impact directly the balance between mRNA and protein levels (Bhaskaran and Mohan 2014).

Since then, one of the most studied regulatory mechanisms that directly mediate mRNA and protein translation are the miRNAs. These non-coding short RNA molecules inhibit the translation and alter the stability of mRNA by binding to complementary sites on the target mRNAs, usually in the 3′ UTR. With such capabilities, miRNAs are responsible for coordinately controlling genes expression involved in several cellular mechanisms, such as inflammation, cell cycle, apoptosis, migration, and stress, among others pathways involved in disease development (Mollaei et al. 2019). Most importantly, miRNA alterations are evident in several cancer types and correlated with differentiation stages. These molecules can be miRNAs tumor suppressor or oncogenic (oncomiRS). For example, in prostate, pancreatic, bladder cancers, and multiple myeloma the tumor suppressor miR-145 controls targets such as ROCK1, p-AKT, p-PI3K, STAT3, and FOXO1 (Kato et al. 2017; Mollaei et al. 2019). On the other hand, in breast cancer the oncomIR controls PTEN/Akt pathway and contributes to tumorigenesis (Kato et al. 2017; Li et al. 2017).

The main action of ncRNAs widely known is the negative regulation of gene expression by binding a target mRNA through complex formation and induction of its degradation or inhibition of its translation by different mechanisms (Ha and Kim 2014). Regulatory ncRNAs can be divided into microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs) (Ozata et al. 2019), small interfering RNAs (siRNAs), and long non-coding RNAs (lncRNAs) (Yao et al. 2019). The largest quantitative contribution to the group of the non-protein-coding transcripts belongs to the group of lncRNAs, which are arbitrarily considered as about 200 nucleotides in length. Since many of these lncRNAs can also act as primary transcripts for the production of short RNAs, they are involved in the silencing of gene expression (Ponting et al. 2009). In summary, these inhibitory molecules provide possible explanations on how variations can arise between transcriptomics and proteomics profiles in biological systems.

8.4.4 Protein Degradation

On the other side of abundance control, protein half-life can vary significantly depending on a number of different conditions (Glickman and Ciechanover 2002). The proteome is modulated by protein degradation rates, which are influenced by protein localization, stability, the three-dimensional conformation, and their integration into stable protein complexes. The amino-terminal and carboxy-terminal composition of a protein can determine a protein's half-life through the recognition of degron sequences by proteolytic systems that cause degradation via N-degron pathways or C-degron pathways, respectively (Qian et al. 2003).

To keep cellular homeostasis, cells evolved a dynamic and self-regulating quality control processes to maintain protein and to prevent accumulation of damaged molecules. Considering that approximately 240 g protein are synthesized and degraded daily in a 60 kg adult human (Mitch and Goldberg 1996), no wonder a failure on this tight turnover system ultimately leads to disease. In cells, protein degradation is achieved by different degradation systems, of which the ubiquitin–proteasome system (UPS) and autophagy are involved in the degradation of the majority of cellular proteins. Yet another function of proteolytic pathways is selective destruction of proteins whose concentrations must vary with time and alterations in the state of a cell.

The UPS mostly degrades single, unfolded polypeptides able to enter into the narrow channel of the proteasome, and the majority of intracellular proteins are degraded by this process (Zhao et al. 2015). UPS comprises the ubiquitylation system, which involves the activity of specific enzymes that ubiquitylate or deubiquitylate target proteins, and the proteasome system, which degrades ubiquitylated proteins (Collins and Goldberg 2017). Ubiquitylation is a sequential, ATP-consuming process involving a hierarchically acting enzymatic cascade E1, E2, and E3 enzymes, which mediate the covalent attachment of ubiquitin monomers (monoubiquitylation) or chains (polyubiquitylation) to protein substrates. Ubiquitin (Ub) is typically attached via its carboxy-terminus to a lysine residue on a target protein, and it contains seven lysine residues, Lys6, Lys11, Lys27, Lys29, Lys33, Lys48, and Lys63, which can form up to seven different polyubiquitin chain linkages. The mode of conjugation determines the fate of ubiquitylated proteins, including targeting proteins for degradation, affecting their activity or altering their localization. The proteasome preferentially degrades branched (Lys48-linked) polyubiquitylated proteins, although chains containing nearly all linkages can be recognized and degraded by the proteasome (Meyer and Rape 2014).

The proteasome is the most complex protease in the UPS, has a molecular mass >2.5 MDa, and exists in multiple structural forms but contains two assemblies, a proteolytic chamber formed by the core particle (20S) and a regulatory particle (19S or PA700), which are functionally linked by a gated protein translocation channel, which collectively are known as the 26S complex. Although the roles of many of 26S subunits and associated proteins are still unclear, the 26S proteasome catalyzes the great majority (at least 80%) of the protein degradation in growing mammalian cells. Of note, the proteasome does not degrade proteins to individual amino acids but instead polypeptides are digested to short peptides, which range between 2 and 10 residues in length. The remaining peptides are digested in seconds to amino acids by cytosolic peptidases, but in mammals some serve as precursors for antigenic peptides displayed on MHC-class I molecules (Kisselev et al. 1999; Murata et al. 2018).

As proteolysis is irreversible, intricate multi-level mechanisms have evolved to ensure efficient and selective protein degradation. In this scenario removal of ubiquitin from substrates is tightly controlled by deubiquitylating enzymes (DUBs; also known as deubiquitylases or deubiquitinases). There are four main families of DUBs, and they cleave ubiquitin from proteins and disassemble polyubiquitin chains that are released from substrates before proteasomal degradation, recycling Ub for subsequent ubiquitylation reactions, preventing proteasome congestion and controlling protein turnover by modifying or removing ubiquitin or polyubiquitin chains from the targeted protein (Wilkinson 1997).

Another important, although less specific protein degradation machinery is called autophagy, which is an intracellular pathway for bulk protein degradation and the removal of damaged organelles by lysosomes. It is involved in recycling cellular components like the cytoplasmic proteins; soluble misfolded protein and insoluble misfolded aggregates content for reuse and ensuring that it obeys the rule that "energy can neither be created nor destroyed instead it can be change from one form to another" as the energy required for degradation is high which is also in tandem with notable energy also biosynthesis (Wang et al. 2015). Thus, it helps to change the state of cellular contents to re-useable form to build new cells. There are four pathways identified for the autophagic process which include: the post-translational modification dependent and independent CMA pathways and the ubiquitin dependent and independent macroautophagy pathways (Wang et al. 2015). Autophagy occupies a central position in the maintenance of cellular homeostasis by directing protein degradation, and the process adapts cells to adverse micro-environmental conditions mainly stress such as nutrient/energy starvation, hypoxia, ER stress, hypoxia, and organelle damage (Chen et al. 2019). A precarious balance is essential in protein synthesis as well as turnover so as to prevent the onset of diseases such as neurodegeneration and cancer, which has made autophagy pathway a target in the management of these diseases (Dikic 2017).

Taken together, these major processes and machineries discussed above bring their individual roles into the complex network of events that keep the cellular homeostasis. In the next sections, we will explore how these individual processes contribute to our understanding about when, during cellular events, it is possible to expect a balanced correlation between RNA and protein levels.

8.5 Temporal Correlation of RNA and Protein Levels

In healthy cells and at steady state, RNA and protein synthesis and degradation are well balanced (Harper and Bennett 2016). A given protein abundance can be obtained from infinitely many combinations of these synthesis and degradation rates. In addition, variation in mRNA abundance is frequently buffered on the protein level, meaning that a substantial change in mRNA abundance is not propagated immediately into a corresponding change in protein abundance (Liu et al. 2016). The cell can control the rates of degradation or synthesis for a given protein, and there is significant heterogeneity even within proteins that have similar functions (Pratt et al. 2002). It is clear that cells control protein production at multiple levels, and the resulting amounts of protein reflect cellular integration of the various regulatory layers, ranging from mRNA production to protein degradation. Although regulation at a single level might prevail in some cases, it is common for cells to coordinately modulate gene expression at several levels.

One important reason for a general lack of correlation between mRNA and protein abundance may be that proteins have very different half-lives as the result of varied protein synthesis and degradation ratios. Given that proteins are on average more stable than mRNAs, proteins can still be present when the mRNA that encoded them is long gone. Therefore, it seems recognizable that the possible reasons why imperfect mRNA–protein correlations arise from the majority of studies for that matter is the factor of "time." Any change in the transcriptional state of a cell will lead to a delay in the response at the protein level simply due to the time it takes to reach a new steady state. Correlations at specific time points during a transition may be uninformative, as changes in mRNA levels in reality correspond to latent changes in protein levels that have yet to occur. Indeed, examples of this are seen in many studies that will be discussed ahead.

As an example, at steady state, the RNA polymerase II machinery can transcribe 2–6 kb/min for a mammalian cell (Maiuri et al. 2011). Considering an average length for an mRNA around 2kb, it takes around a couple of minutes to transcribe one gene. On the protein side, the ribosomal machinery operates at a rate of few to several amino acids per second, generating proteins with average length also in a couple of minutes. Moreover, since many ribosomes can translate the same molecule of mRNA simultaneously, these rates can increase significantly (Riba et al. 2019). Overall, it takes more than an hour to generate 10^6 protein molecules after initiation of transcription from a single locus, for example during cell duplication. A faster means to upregulate proteins is to increase the number of mRNA molecules, amplifying their translation exponentially (Schwanhäusser et al. 2013). However, many factors, such as rates of translation initiation, sequence, folding, and structure of the protein, also significantly affect these rates, again disrupting the correlation of mRNA and protein levels (Riba et al. 2019).

8.6 The Imbalance Between the Transcriptome and Proteomes: Lessons Learned from High-Throughput Studies

Considering what has been discussed so far, it is certain that protein and RNA-based measurements are complementary to provide accurate status of cellular homeostasis. It is important to recognize that many factors can cause imbalances between levels of messenger (transcript) and its final effector (mature protein). As we have seen, several post-transcriptional and post-translational control mechanisms such as the translation rate or half-lives of mRNAs and proteins are affected by a wide range of factors.

On top of all the biological dynamics, the methods of RNA sequencing and also of protein expression evaluation by mass spectrometry suffer from technical limitations that affect the precision and final accuracy of the quantitative measurements. For example, biases in RNA data can arise during formation of the sequencing library. Also, in mass spectrometry, the shotgun approach relies on proteins digested with enzymes such as trypsin to generate peptides that are the entities identified. The tremendous variety of chemical species generated in shotgun approaches, which are influenced by several physicochemical factors, and the stochastic property of this technique, turn the proteome samples extremely challenging and almost impossible to be completely characterized. Certain approaches such as the use of isotopic labeling for relative quantitation or the run of multiple technical and biological replicates in mass spectrometry turn such proteome complexity into a feasible strategy that can be effectively profiled (Buccitelli and Selbach 2020).

Even with the significant developments in the technologies used to quantify protein abundance over the past couple of years, protein identification and quantification still lag behind the high-throughput experimental techniques used to determine mRNA expression levels. The proteome of a cell or tissue at a specific time point is extremely complex and diverse. The major limits of proteome analysis are associated with the heterogeneity of proteins and the huge differences in abundance (dynamic range). Abundant proteins mask the presence of low abundant proteins. Because no PCR equivalent exists for proteins, low-abundant proteins have a low probability to be detected (Churchill 2002; Larkin et al. 2005). Since it is fundamental to consider at a least some of these drawbacks reported above for a satisfactory comparison between the transcriptome and proteome, several studies have been specifically designed with this particular focus.

Overall, genome-wide studies have shown that the correlation between expression levels of mRNA and protein are marginal, hovering around 40–50% across many studies. One of the seminal studies specifically developed with the purpose to compare the expression profile of active genes in the adult human liver and the protein abundance in human plasma (Kawamoto et al. 1996). The study found a positive correlation between the abundance of the transcript and the protein concentration in the serum. It was also possible to categorize the responsible genes into three groups: those with less than five transcripts (per 1000 mRNA molecules) produce proteins at a level of <0.1 g/100ml, those with 5–20 transcripts produce proteins at 0.1–0.4 g/100 ml, and those with more than 30 transcripts produce proteins at 0.5–4 g/100 ml. This was a pioneering study on a large scale showing that particularly for secreted proteins, the transcript – protein correlation was positive.

Another important study explored for the first time a quantitative comparison of mRNA transcript and protein expression levels for a relatively large number of genes expressed in the same metabolic state in yeast (Gygi et al. 1999). The study concluded that predictions of protein levels from mRNA transcript levels were not feasible. This study particularly relied on 2D-electrophoresis to evaluate the proteome, which itself is a very limited technique in terms of dynamic range for protein abundance. However, the study found that for a subset of 106 highly abundant proteins, the correlation with mRNA levels was positive. Subsequently, using the same yeast as model system, it was again demonstrated a partial correlation of protein expression after specific perturbations in know pathways, namely the galactose utilization (Ideker et al. 2001). While several genes–proteins ratios correlated well in increased or decreased expression upon perturbation, others know players in the pathway still had poor correlation. More specifically this study attributed the discrepancy in gene–protein expression correlations to post-transcriptional regulatory events. Yet, this study also uncovers that for genes linked by physical interactions in the network tend to have more strongly correlated expression profiles than genes chosen at random. Using modern high-throughput proteomics and accurate relative quantitation based on stable isotopes, another study explored in depth the yeast proteome and transcriptome correlation (de Godoy et al. 2008). This study once again demonstrated the poor overall correlation of transcriptome and proteome, but particularly found that good correlation was found for the subset of genes involved in yeast pheromone pathway components.

Studies based on more complex organisms also provided contrasting mRNA and protein levels. In a detailed comparison of mesenchymal stromal cells obtained from bone marrow or umbilical cord vein, with the overall objective to prove the interchangeability of these sources for cellular therapy, proteomic and gene expression analysis reached a 63% correlation level for those specific set of genes specific for one or the other mesenchymal cell type (Miranda et al. 2012). This dataset is particularly illustrated in the correlation plot (Fig. 8.2a), which indicated that mRNA abundance data (y-axis) presented more spreading in terms of ratios in comparison to proteomic data (x-axis). Using more sophisticated proteomic strategies, the dynamics of protein and mRNA expression levels across the cell cycle in human myeloid leukemia cells using was explored (Ly et al. 2014). Myeloid-specific gene expression and variations in protein abundance, isoform expression, and phosphorylation at different cell cycle stages were dissected for over ~6000 genes individually across the cell cycle, revealing complex, gene-specific patterns. Protein and mRNA correlations were modest across different cell cycle stages, suggesting again greater contribution of post-transcriptional mechanisms in cell cycle control.

Considering that most of the aforementioned studies focused on static of minimally dynamic biological events, the temporal contribution to the lack of correlation between proteome and transcriptome of a cell was still obscure. A breakthrough study shed light in the time variable studying the dynamics of embryonic development (Peshkin et al. 2015). Based on a time-resolved deep quantitative profiling of proteins and mRNA, the study produced an unprecedented dataset that illustrated the turnover of these molecules during the embryo development. As example, Fig. 8.2b demonstrates the normalized curves of mRNA and protein levels over time. It is clear that the initial wave of mRNA expression and accumulation, was followed by a quick decay, while protein levels progressively accumulated for both CAPN8 and LIN28A genes. Obviously, depending on the moment one makes the mRNA and protein measurements for such gene, more or less correlation will be found. On the other hand, the other illustrated genes, DND1 and SPARC, have a

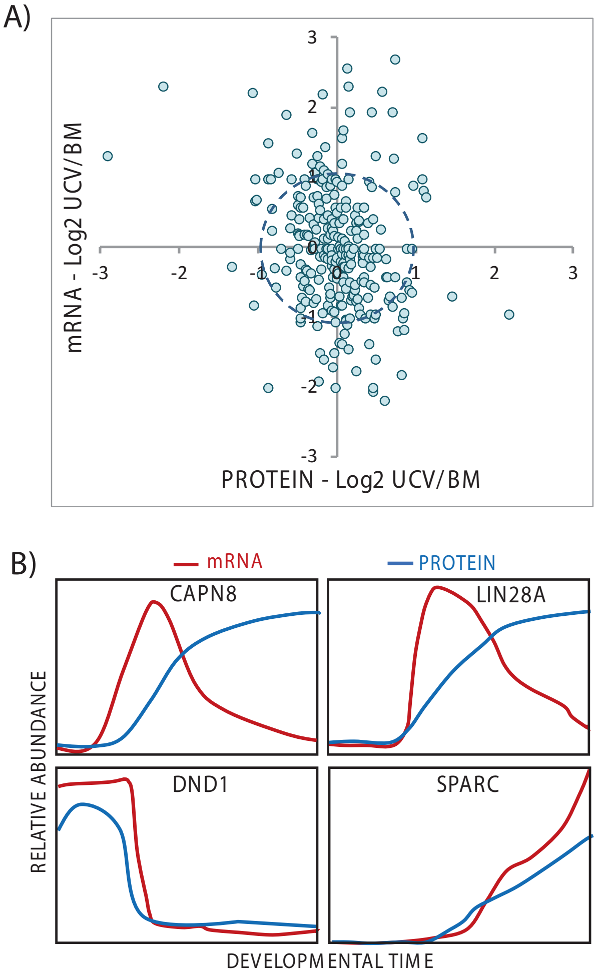


Fig. 8.2 Correlation of gene expression in complex systems. (**a**) The combined analysis of the transcriptome and proteome of mesenchymal stromal cells from bone-marrow (BM and umbilical cord vein (UVC) demonstrated a correlation of 63% of the profiled genes (central circle) and similarity between these two sources of therapeutic cells (reproduced from Miranda et al. 2012). (**b**) Time-course experiments during embryonic development demonstrated the syntheses and decay of mRNA and proteins. While some genes (CAPN8 and LIN28A) have an evident difference in timing for synthesis and degradation, others (DND1 and SPARC) present a completely synchronized and correlated gene expression. (Reproduced from Peshkin et al. 2015)

very tight correlation. Of note, DND1 itself is a RNA-binding factor that positively regulates gene expression by prohibiting miRNA-mediated gene suppression, creating a scape for post-transcriptional regulation. SPARC is a secreted protein, class that has been observed with greater gene expression correlations.

Several other studies support these major findings described above. But particularly for diseases, the context of protein versus mRNA expression becomes particularly important for diagnostics and molecular profiling. As examples, a study correlated the expression of microRNA, mRNA, and proteins in the identification of microRNA-related cancers, particularly in glioblastoma (Seo et al. 2017). For a subset of 146 upregulated genes, mRNA and proteins were positively correlated. These findings are consistent with the hypothesis that the malignant phenotype required additional cancer promoter genes that were coordinately overexpressed. In a similar study, Yang and colleagues used a combination of proteomics and transcriptomics strategies and found potential targets in early colorectal cancer (CRC) (Yang et al. 2019). The study identified 2968 proteins in stage II CRC proteomics data, where most (2846) of these proteins were identified in TGCA transcriptome data. Numerous bioinformatics methods, including differential expression analysis, weighted correlation network analysis, gene ontology, and protein–protein interaction analyses, were used to select a set of 111 key proteins, differentially expressed in terms of proteins and mRNAs, levels. These highly correlated genes can represent a molecular signature for the CRC, and used, for example, to subclassify the tumor types. In summary, the integration of proteomics and transcriptomics data, particular for disease studies, can generate a high-resolution global expression map that can collaborate to discover new biomarkers for several diseases.

8.7 Final Remarks and Perspectives

As we discussed in this chapter, profiling gene expression enables a global physiological picture for a given system in a specific context or moment. When the dynamics of cellular processes is taken into account, several regulatory processes emerge and explain apparent disconnection of the transcriptome and proteome. Unlike the genome, which is virtually static in terms of its composition and size, we gave several examples here that support the dynamics of the genetic cellular programming, which continually changes depending on the phase of the cell cycle, the organ, exposure to drugs or physical agents, aging, diseases such as cancer and autoimmune diseases, and a multitude of other variables.

Several of the factors that modulate abundance of mRNA and proteins have been presented. New features of these molecular mechanisms have been continuously uncovered, mainly promoted by advances in high-throughput deep biomolecular profiling. In addition to the development of modern multidimensional transcriptomics and proteomics strategies, bioinformatics and data integration have become a common basis for translational areas, where complex integrated mRNA and protein molecular signatures are aiding the development of new therapeutic strategies or methods for diseases diagnostics. Ultimately, full and effective integration across the relatively static genomic information with the dynamic transcriptomic and proteomic data will produce complete maps of normal and pathological process to drive personalized medicine.

With the continuous advancing of technology and biology, the interplay of transcriptomics and proteomics profiles in living organisms will become more evident and fundamental to provide answers to many relevant biological questions in health and disease.

Support The authors of this work were supported by grants from the following Brazilian funding agencies: FAPESP; J.O.T -(grant no. 2020/13463-2), CNPq, and CAPES.

References

Anderson L, Seilhamer J (1997) A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18:533–537

Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH (2017) Proteomics: technologies and their applications. J Chromatogr Sci 55:182–196

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJ (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7:246

Barbulovic-Nad I, Lucente M, Sun Y, Zhang M, Wheeler AR, Bussmann M (2006) Bio-microarray fabrication techniques—a review. Crit Rev Biotechnol 26:237–259

Beilharz TH, Preiss T (2004) Translational profiling: the genome-wide measure of the nascent proteome. Brief Funct Genomic Proteomic 3:103–111

Bernstein JA, Khodursky AB, Lin P-H, Lin-Chao S, Cohen SN (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. PNAS 99:9697–9702

Bhaskaran M, Mohan M (2014) MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. Vet Pathol 51:759–774

Brenner S, Jacob F, Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature 190:576–581

Buccitelli C, Selbach M (2020) mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet 21:630–644

Buttgereit F, Brand MD (1995) A hierarchy of ATP-consuming processes in mammalian cells. Biochem J 312:163–167

Chandramouli K, Qian P-Y (2009) Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. Hum Genomics Proteomics 2009: 239204

Chen R-H, Chen Y-H, Huang T-Y (2019) Ubiquitin-mediated regulation of autophagy. J Biomed Sci 26:80

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet 32:490–495

Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, Xiao W, Laudanski K, Brownstein BH, Elson CM, Hayden DL, Herndon DN, Lowry SF, Maier RV, Schoenfeld DA, Moldawer LL, Davis RW, Tompkins RG, Program§§ I and HR to IL-SCR (2005) Application of genome-wide expression analysis to human health and disease. PNAS 102:4801–4806

Collins GA, Goldberg AL (2017) The logic of the 26S proteasome. Cell 169:792–806

Darzacq X, Shav-Tal Y, de Turris V, Brody Y, Shenoy SM, Phair RD, Singer RH (2007) In vivo dynamics of RNA polymerase II transcription. Nat Struct Mol Biol 14:796–806

de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, Walther TC, Mann M (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455:1251–1254

Dever TE (2002) Gene-specific regulation by general translation factors. Cell 108:545–556

Dikic I (2017) Proteasomal and autophagic degradation systems. Annu Rev Biochem 86:193–224

Dunham I, Kundaje A, Aldred SF, et al, The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Faça VM (2017) Selective reaction monitoring for quantitation of cellular proteins. Methods Mol Biol 1546:213–221

Gebauer F, Hentze MW (2004) Molecular mechanisms of translational control. Nat Rev Mol Cell Biol 5:827–835

Glickman MH, Ciechanover A (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. Physiol Rev 82:373–428

Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR (2004) Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. Mol Cell Biol 24:5534–5547

Gutiérrez R, Ewing R, Cherry J, Green P (2002) Identification of unstable transcripts in Arabidopsis by cDNA microarray analysis: rapid decay is associated with a group of touch- and specific clock-controlled genes. Proc Natl Acad Sci USA 99: 11513-11518

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730

Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15:509–524

Harper JW, Bennett EJ (2016) Proteome complexity and the forces that drive proteome imbalance. Nature 537:328–338

Hausser J, Mayo A, Keren L, Alon U (2019) Central dogma rates and the trade-off between precision and economy in gene expression. Nat Commun 10:68

Heber S, Sick B (2006) Quality assessment of affymetrix GeneChip data. OMICS J Integr Biol 10:358–368

Hegde PS, White IR, Debouck C (2003) Interplay of transcriptomics and proteomics. Curr Opin Biotechnol 14:647–651

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–934

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martínez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W (2005) Multiple-laboratory comparison of microarray platforms. Nat Methods 2:345–350

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356

Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, Zhang Z, Harland RM (2015) Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. Cell Mol Life Sci 72:3425–3439

Kato M, Kurozumi A, Goto Y, Nohata N, Arai T, Okato A, Koshizuka K, Kojima S, Ichikawa T, Seki N (2017) Abstract 1459: dual-strand tumor-suppressor microRNA-145 (miR-145-5p and miR-145-3p) are involved in castration-resistant prostate cancer pathogenesis. Cancer Res 77:1459–1459

Kawamoto S, Matsumoto Y, Mizuno K, Okubo K, Matsubara K (1996) Expression profiles of active genes in human and mouse livers. Gene 174:151–158

Kisselev AF, Akopian TN, Woo KM, Goldberg AL (1999) The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes: implications for understanding the degradative mechanism and antigen presentation*. J Biol Chem 274:3363–3371

Klose J (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. Hum Genet 26:231–243

Kukurba KR, Montgomery SB (2015) RNA sequencing and analysis. Cold Spring Harb Protoc 2015:pdb.top084970

Lanfredi GP, Thomé CH, Ferreira GA, Silvestrini VC, Masson AP, Vargas AP, Grassi ML, Poersch A, Candido dos Reis FJ, Faça VM (2021) Analysis of ovarian cancer cell secretome during epithelial to mesenchymal transition reveals a protein signature associated with advanced stages of ovarian tumors. Biochimica et Biophysica Acta (BBA) – Proteins Proteomics 1869:140623

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, DG MA, Lek M, Lizano E, HPJ B, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, MI MC, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo Á, Antonarakis SE, Häsler R, Syvänen A-C, van Ommen G-J, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501:506–511

Larance M, Lamond AI (2015) Multidimensional proteomics for cell biology. Nat Rev Mol Cell Biol 16:269–280

Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J (2005) Independence and reproducibility across microarray platforms. Nat Methods 2:337–344

Lee TI, Young RA (2013) Transcriptional regulation and its misregulation in disease. Cell 152:1237–1251

Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854

Li N, Miao Y, Shan Y, Liu B, Li Y, Zhao L, Jia L (2017) MiR-106b and miR-93 regulate cell progression by suppression of PTEN via PI3K/Akt pathway in breast cancer. Cell Death Dis 8:e2796–e2796

Liu Y, Beyer A, Aebersold R (2016) On the dependency of cellular protein levels on mRNA abundance. Cell 165:535–550

Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T (2017) Transcriptomics technologies. PLoS Comput Biol 13:e1005457

Ly T, Ahmad Y, Shlien A, Soroka D, Mills A, Emanuele MJ, Stratton MR, Lamond AI (2014) A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. elife 3:e0163

Maiuri P, Knezevich A, De Marco A, Mazza D, Kula A, McNally JG, Marcello A (2011) Fast transcription rates of RNA polymerase II in human cells. EMBO Rep 12:1280–1285

Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, Ferrari R (2018) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform 19:286–302

Meyer H-J, Rape M (2014) Enhanced protein degradation by branched ubiquitin chains. Cell 157:910–921

Miranda HC, Herai RH, Thomé CH, Gomes GG, Panepucci RA, Orellana MD, Covas DT, Muotri AR, Greene LJ, Faça VM (2012) A quantitative proteomic and transcriptomic comparison of human mesenchymal stem cells from bone marrow and umbilical cord vein. Proteomics 12:2607–2617

Mitch WE, Goldberg AL (1996) Mechanisms of muscle wasting. The role of the ubiquitin-proteasome pathway. N Engl J Med 335:1897–1905

Mollaei H, Safaralizadeh R, Rostami Z (2019) MicroRNA replacement therapy in cancer. J Cell Physiol 234:12369–12384

Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10:135–151

Murata S, Takahama Y, Kasahara M, Tanaka K (2018) The immunoproteasome and thymoproteasome: functions, evolution and human disease. Nat Immunol 19:923–931

Nagalakshmi U, Waern K, Snyder M (2010) RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol Chapter 4:Unit 4.11.1–13.

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. J Biol Chem 250:4007–4021

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD (2019) PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet 20:89–108

Parker R, Song H (2004) The enzymes and control of eukaryotic mRNA turnover. Nat Struct Mol Biol 11:121–127

Peshkin L, Wühr M, Pearl E, Haas W, Freeman RM, Gerhart JC, Klein AM, Horb M, Gygi SP, Kirschner MW (2015) On the relationship of protein and mRNA dynamics in vertebrate embryonic development. Dev Cell 35:383–394

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–641

Pratt JM, Petty J, Riba-Garcia I, Robertson DHL, Gaskell SJ, Oliver SG, Beynon RJ (2002) Dynamics of protein turnover, a missing dimension in proteomics. Mol Cell Proteomics 1:579–591

Preiss T, Hentze MW (2003) Starting the protein synthesis machine: eukaryotic translation initiation. BioEssays 25:1201–1211

Qian J, Kluger Y, Yu H, Gerstein M (2003) Identification and correction of spurious spatial correlations in microarray data. BioTechniques 35:42–48

Raghavan A, Ogilvie RL, Reilly C, Abelson ML, Raghavan S, Vasdewani J, Krathwohl M, Bohjanen PR (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. Nucleic Acids Res 30:5529–5538

Riba A, Nanni ND, Mittal N, Arhné E, Schmidt A, Zavolan M (2019) Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. PNAS 116:15023–15032

Roux PP, Topisirovic I (2018) Signaling pathways involved in the regulation of mRNA translation. Mol Cell Biol 38: e00070-18

Schoenberg DR, Maquat LE (2012) Regulation of cytoplasmic mRNA decay. Nat Rev Genet 13:246–259

Schwanhäusser B, Wolf J, Selbach M, Busse D (2013) Synthesis and degradation jointly determine the responsiveness of the cellular proteome. BioEssays 35:597–601

Selinger DW, Saxena RM, Cheung KJ, Church GM, Rosenow C (2003) Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. Genome Res 13:216–223

Seo J, Jin D, Choi C-H, Lee H (2017) Integration of MicroRNA, mRNA, and protein expression data for the identification of cancer-related MicroRNAs. PLoS One 12:e0168412

Shaw RJ, Cantley LC (2006) Ras, PI(3)K and mTOR signalling controls tumour cell growth. Nature 441:424–430

Silvestrini VC, Lanfredi GP, Masson AP, Poersch A, Ferreira GA, Thomé CH, Faça VM (2019) A proteomics outlook towards the elucidation of epithelial–mesenchymal transition molecular events. Mol Omics 15:316–330

Silvestrini VC, Thomé CH, Albuquerque D, de Souza PC, Ferreira GA, Lanfredi GP, Masson AP, Delsin LEA, Ferreira FU, de Souza FC, de Godoy LMF, Aquino A, Carrilho E, Panepucci RA, Covas DT, Faça VM (2020) Proteomics analysis reveals the role of ubiquitin specific protease (USP47) in Epithelial to Mesenchymal Transition (EMT) induced by TGFβ2 in breast cells. J Proteome 219:103734

Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 13:227–232

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO (2002) Precision and functional specificity in mRNA decay. PNAS 99:5860–5865

Wang D, Peng Z, Ren G, Wang G (2015) The different roles of selective autophagic protein degradation in mammalian cells. Oncotarget 6:37098–37116

Wilkinson KD (1997) Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. FASEB J 11:1245–1256

Wilusz CJ, Wilusz J (2004) Bringing the role of mRNA decay in the control of gene expression into focus. Trends Genet 20:491–497

Wilusz CJ, Wormington M, Peltz SW (2001) The cap-to-tail guide to mRNA turnover. Nat Rev Mol Cell Biol 2:237–246

Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res 13:1863–1872

Yang W, Shi J, Zhou Y, Liu T, Zhan F, Zhang K, Liu N (2019) Integrating proteomics and transcriptomics for the identification of potential targets in early colorectal cancer. Int J Oncol 55:439–450

Yao R-W, Wang Y, Chen L-L (2019) Cellular functions of long noncoding RNAs. Nat Cell Biol 21:542–551

Zhang Y, Kwok-Shing Ng P, Kucherlapati M, Chen F, Liu Y, Tsang YH, de Velasco G, Jeong KJ, Akbani R, Hadjipanayis A, Pantazi A, Bristow CA, Lee E, Mahadeshwar HS, Tang J, Zhang J, Yang L, Seth S, Lee S, Ren X, Song X, Sun H, Seidman J, Luquette LJ, Xi R, Chin L, Protopopov A, Westbrook TF, Shelley CS, Choueiri TK, Ittmann M, Van Waes C, Weinstein JN, Liang H, Henske EP, Godwin AK, Park PJ, Kucherlapati R, Scott KL, Mills GB, Kwiatkowski DJ, Creighton CJ (2017) A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations. Cancer Cell 31:820–832.e3

Zhao J, Zhai B, Gygi SP, Goldberg AL (2015) mTOR inhibition activates overall protein degradation by the ubiquitin proteasome system as well as by autophagy. PNAS 112:15790–15797

Chapter 9
Transcriptome During Normal Cell Differentiation

Karina Fittipaldi Bombonato-Prado, Adalberto Luiz Rosa, Paulo Tambasco de Oliveira, Janaína Andrea Dernowsek, Vanessa Fontana, Adriane Feijó Evangelista, and Geraldo A. Passos

9.1 Human Mesenchymal Stem Cells Represent a Model-System for Cell Differentiation Studies

Stem cells are mainly classified as adult stem cells (ASCs), embryonic stem cells (ESCs), and induced pluripotent stem cells (iPSCs), in which mesenchymal stem cells (MSCs) are considered the main class of ASCs with prominent therapeutic efficacies (Dayem et al. 2019). The progressive restriction of the differentiation potential from pluripotent embryonic stem cells (ESC) to different populations of adult stem cells depends on the orchestrated action of key transcription factors and changes in the profile of epigenetic modifications that ultimately lead to the expression of different sets of genes. ESC are unique in their capacities to self-renew and differentiate into any somatic and germline tissue, while, by contrast, the differentiation potential of adult stem cells is limited (Aranda et al. 2009).

K. F. Bombonato-Prado (✉) · P. T. de Oliveira · G. A. Passos
Department of Basic and Oral Biology, School of Dentistry of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brazil
e-mail: karina@forp.usp.br

A. L. Rosa
Department of Oral and Maxillofacial Surgery and Periodontology, School of Dentistry of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brazil

J. A. Dernowsek
BioEdTech, São Paulo, SP, Brazil

V. Fontana
Department of Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, UK

A. F. Evangelista
Pio XII Foundation, Barretos Cancer Hospital, Barretos, SP, Brazil

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_9

Studies have shown that mesenchymal stem cells (MSCs) reflect the stem cell differentiation potential and may form the basis of studies designed to provide insights into genes that confer the greatest developmental potency (Ulloa-Montoya et al. 2007). The knowledge of the fundamental processes associated with the differentiation of MSCs is still poor, and elucidation of the genetic cascade guiding these cells to become more specialized is important for both basic knowledge and clinical application (de Jeong et al. 2004). Roson-Burgo et al. (2016) described a core mesenchymal lineage signature of 489 genes based on a deep comparative analysis of multiple transcriptomic expression data series that comprise MSCs of different tissue origins, MSCs undifferentiated states of commitment and other related non-mesenchymal human cell types.

MSCs have now been isolated from many sites throughout the body. In the bone compartment, they can be found in bone marrow, periosteum, endosteum, and bone mineralized matrix itself, and are known to be the primary sources of cells during bone repair (Knight and Hankenson 2013).

The differentiation of MSCs toward different lineages seems to display different metabolism signatures (Chen et al. 2014). For instance, there is a transition from glycolysis to oxidative phosphorylation in MSCs' differentiation toward osteogenic lineage (Chen et al. 2008) and adipogenic lineage (Hofmann et al. 2012; Tormos et al. 2011). In contrast, when MSCs differentiate toward chondrogenic lineage using pellet culture, glycolysis is enhanced (Pattappa et al. 2011). Furthermore, it has been reported that mitochondrial metabolism and reactive oxygen species (ROS) generation might be one of the causal factors rather than the merely results of adipogenic differentiation (Tormos et al. 2011). Thus, treatments altering mitochondrial metabolism and ROS generation might affect or determine MSCs' fate.

Several methods for MSCs' generation have been developed, relying initially on cell scraping, followed by trypsinization, defined culture conditions, and more recently the utilization of three-dimensional (3D) platforms for MSCs derived via spheroid culture (Dayem et al. 2019). Former studies indicate that MSCs originating from specific tissues are capable of differentiation into distinct tissues (Kim et al. 2006). The molecular characteristics, surface antigen expression, and biological functions such as proliferation and differentiation capacities of MSCs can vary based on the MSC source (Billing et al. 2016). For instance, bone marrow MSCs presented higher expression of genes related to osteogenesis, whereas adipose tissue MSCs showed a higher expression of genes related to angiogenesis and adipocyte differentiation, irrespective of cell differentiation (Fidelis et al. 2019). Such differences lead the authors to suggest that the former should be considered for bone regeneration and adipose tissue MSCs for angiogenesis (Fidelis et al. 2019). In addition to the bone marrow (BM), MSCs have been found in several other sites such as circulating blood of preterm fetuses, hematopoietic lineage (Campagnoli et al. 2001; Erices et al. 2000), Wharton's jelly explants (Ishige et al. 2009; Wagner et al. 2005), adipose tissue (Xu et al. 2017), oral tissues (Zhou et al. 2020), and lung and dermal tissues (Yaghoubi et al. 2019). According to Souza et al. (2016), the perivascular MSCs are adventitial cells, acting as precursors of the pericytes (Yianni and Sharpe 2019) and other stromal cells during tissue homeostasis.

Although the presence of MSCs in the umbilical cord vein (UC) of newborns was formerly controversial (Mareschi et al. 2001; Wexler et al. 2003), this site is now being used as a standard source of these cells. Sarugaser et al. (2005) have shown that perivascular tissue from human UC vein cultivated in a non-osteogenic medium contains a subpopulation of cells with an osteogenic phenotype that forms calcified nodules. The addition of osteogenic chemical supplementation to the culture medium resulted in a significant increase of these cells. Wang et al. (2004) demonstrated that mesenchymal cells from the mucous connective tissue of Wharton's jelly express matrix receptors (CD44, CD105) and integrin markers (CD29, CD51), suggesting that these cells are similar to stem cells in that they can be differentiated into chondrogenic, adipogenic, or osteogenic cell lines. Recent findings showed that exosomes produced by mesenchymal stem cells from human umbilical cord carry biomolecules that promote similar functions to MSCs with low immunogenicity and no tumorization, and therefore play an important role in cell–cell communication (Yaghoubi et al. 2019).

9.2 Therapeutic Potential of Human Mesenchymal Stem Cells

The therapeutic potential of stem cells is already a reality but there is still a need of understanding several aspects of their molecular biology during differentiation and induced pluripotency (Cohen and Melton 2011; Zhao et al. (2021). Stadtfeld and Hockedlinger 2010). Zhao et al. (2021) suggest that different populations of resident stem cells are mobilized at different times and during disease to generate precursors for cell differentiation, providing insight for novel therapeutic approaches. Considering that the control of messenger RNA (mRNA) transcription corresponds to the first step of gene regulation (Rajewsky 2011), which ultimately controls the process of differentiation, transcriptome analysis is critical for better understanding MSCs. The gene expression of pluripotency-related genes has been examined in MSCs derived from bone marrow, adipocytes, amniotic membrane and epithelial endometrium-derived stem cells, and stroma endometrium-derived stem cells, and these studies suggest that pluripotency-related gene expression varies in different tissues (Tanabe 2014). Sacchetti et al. (2016) showed that human cell populations from different anatomical sources, regarded as MSCs, differ widely in their transcriptomic signature and in vivo differentiation potential, but share the capacity to guide the assembly of functional microvessels in vivo, regardless of their anatomical source, or in situ identity as pericytes or circulating cells.

The knowledge of distinct gene modulation is being applied in the investigation of diseases or responses to damage or trauma. Babb et al. (2017) indicated the importance of molecular events that initiate MSCs to proliferate and differentiate in response to damage, showing that Axin2-expressing cells act as their source of Wnt ligands to induce repair via autocrine Wnt/β-catenin signaling. Semeghini et al.

(2018) have observed that the expression of mRNAs and miRNAs in cells from different sites, e.g., bone marrow and calvaria, was distinctively modulated in healthy and osteoporotic rats, suggesting that osteoporosis promotes specific gene expression of osteoblastic cells depending on its site of origin. Investigating the genes that might act as triggers of MSC early differentiation regardless of their tissue of origin is a promising approach, and we hypothesized that MSCs isolated from different anatomical sites (bone marrow and umbilical cord vein) stimulated to differentiate toward a specific cell type would express a set of common genes implicated in the differentiation fate.

9.3 Transcriptome Analysis During Mesenchymal Stem Cell Differentiation

To explore a larger set of genes (transcriptome profiling) in MSC obtained from human bone marrow (BM) and umbilical cord vein (UCV), (Figs. 9.1 and 9.2), we used microarray screening. As expected, the results showed that during early differentiation, BM and UCV cells expressed exclusive sets of genes. However, these two isolates shared expression of 25 genes (Table. 9.1), including those involved in cell–substrate junction assembly/cell–cell adhesion mediated by integrin (integrin, alpha 5, fibronectin receptor, ITGA5), hormone-mediated signaling pathway/ossification (thyroid hormone receptor alpha, THRA), cell differentiation (nephronectin, NPNT), and regulation of cell growth (HtrA1 serine peptidase 1, HTRA1). Based on their involvement with the molecular/biological processes mentioned above, these could be considered key genes in driving early osteoblastic differentiation of MSCs, independent of their anatomic origin.

Earlier studies have compared the gene expression profile of BM stem cells, UCV cells, and other types of stem cells using serial analysis of gene expression (SAGE) (Panepucci et al. 2004), real-time PCR (Guillot et al. 2008), and microarrays (Bombonato-Prado et al. 2009; Carinci et al. 2004; Jeong et al. 2005; Schilling 2008; Shi et al. 2001; Secco et al. 2009) following an extended culture of the MSCs in osteogenic medium. Of note, the previous results of Kulterer et al. (2007) have revealed the participation of the genes ID4, CRYAB, and SORT1 that were considered to be candidates as regulators of osteogenic differentiation.

In this investigation, we hypothesized that during the initial stages, as early as 24 to 168 h into in vitro cultivation, key genes are activated during the critical period in which the fate of MSCs is defined toward osteogenic differentiation, independent of their anatomical origin. A set of 115 specific genes were found in bone marrow MSCs, from which we highlight selected genes including Biglycan (BGN), whose coded protein is a proteoglycan of the extracellular matrix that is involved in the

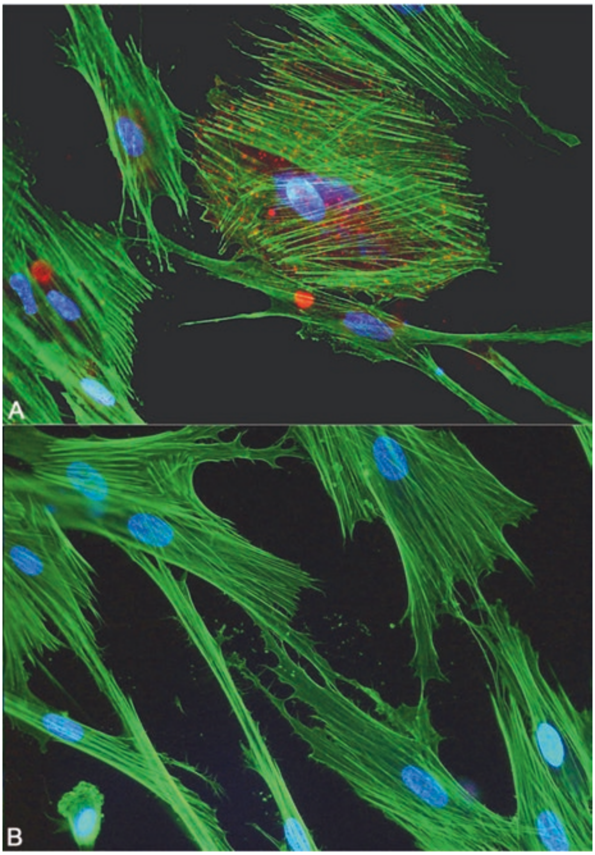


Fig. 9.1 Cell morphology and ALP expression in human umbilical cord stem cells after 7 days of culture: (**a**) cells in contact with osteogenic medium, showing polygonal shape and expression of ALP; (**b**) cells in the absence of osteogenic medium. Green labeling shows actin and blue stain labels cell nuclei (DAPI). Magnification of 400×, fluorescence microscopy

adhesion of collagen fibers (SOURCE Database). This protein is an extracellular matrix structural constituent, which may be involved in collagen fiber assembly (by similarity). Inkson et al. (2009) suggested that WNT1 inducible signaling pathway protein 1 (WISP-1) and BGN may functionally interact and control each other's activities, thus regulating the differentiation and proliferation of osteogenic cells. Besides playing a crucial role in osteogenesis (Jongwattanapisan et al. 2018), BGN expression can be inhibited by microRNAs (miRs) such as *miR-185* during osteoblast differentiation (Cui et al. 2019).

Another modulated gene was fibronectin (FN), which codes the fibronectin protein that binds cell surfaces and various compounds including collagen, fibrin, heparin, and actin (SOURCE Database). Fibronectins are involved in cell adhesion, cell motility, opsonization, wound healing, and maintenance of cell shape. Ogura et al. (2004) also found that MSCs can differentiate into osteoblasts and that FN can stimulate the attachment and spreading of these cells.

The collagen, type VI, alpha 3 (COL6A3) gene codes the alpha-3 chain, one of the three alpha chains of type VI collagen, beaded filament collagen found in most connective tissues. The alpha-3 chain of type VI collagen is much larger than the alpha-1 and -2 chains. These domains have been shown to bind extracellular matrix proteins, an interaction that explains the importance of this collagen in organizing matrix components (SOURCE Database).

A set of 178 umbilical cord-specific genes was modulated, including Tafazzin (TAZ), which codes the tafazzin protein. Tafazzins compose a group of proteins that promote the differentiation and maturation of osteoblasts while preventing adipocyte maturation (SOURCE Database). A large number of morphogens, signaling molecules, and transcriptional regulators have been implicated in regulating bone development, including transcriptional factors like TAZ, Runx2, Osterix, ATF4, and NFATc1 and the Wnt/beta-catenin, TGF-beta/BMP, FGF, Notch, and Hedgehog signaling pathways (Burns et al. 2010; Deng et al. 2008). Recent studies have observed that TAZ may serve as a decisive factor involved in the osteogenesis also in human BM stromal cells when associated with polydatin, a process mediated through the BMP2-Wnt/β-catenin signaling pathway (Shen et al. 2020). Another modulated gene is the microfibrillar-associated protein 3 (MFAP3), which codes a microfibrillar protein important for the structure of extracellular matrix (Abrams et al. 1995), the expression of which has been correlated with bone formation (Burns et al. 2010).

The Sprouty 2 gene (SPRY2) codes a protein associated with cell signaling and cell fate commitment and also plays a role as a modulator of FGF signaling. Welsh et al. (2007) demonstrated that mice carrying a deletion that removes the FGF signaling antagonist Spry2 showed cleft palate, suggesting a role for this gene in the differentiation of MSCs into osteoblasts. Moreover, it was observed that this protein modulates tyrosine kinase signaling, regulating cell migration and proliferation (Edwin et al. 2008). More recently, Vesela et al. (2019) showed decreased bone formation in postnatal Spry2−/− mice, demonstrating the impact of Spry2 deletion in bone biology that included effect on osteoblasts (Runx2) and osteocytes (Sost).

These transcriptional profiles obtained with monolayer cultures are comparable to those obtained with MSCs cultured in three-dimensional scaffolds (Burns et al. 2010), which mimic the in vivo bone formation. This demonstrates that the monolayer culture model-system reproduces the transcriptional modulation of three-dimensional cultures, at least for the genes above mentioned and therefore is adequate to study gene profiling of human MSCs' differentiation.

Finally, we found 25 differentially expressed genes (Figs. 9.3 and 9.4) that were shared between the two MSC sources. Due to the biological processes in which these genes participate, they can be considered triggers of osteoblastic differentiation of MSCs independent of their anatomical origin. Among these, we will discuss selected genes. The integrin, alpha 5 (fibronectin receptor, alpha polypeptide) gene (ITGA5), which is associated with cell-matrix adhesion, was also one of these 25 genes. Integrins are cell surface receptors that interact with the extracellular matrix (ECM) and mediate various intracellular signals, defining cellular shape and mobility and regulating the cell cycle (SOURCE Database). Integrins may play significant roles in determining osteoblast function because they are signal transduction

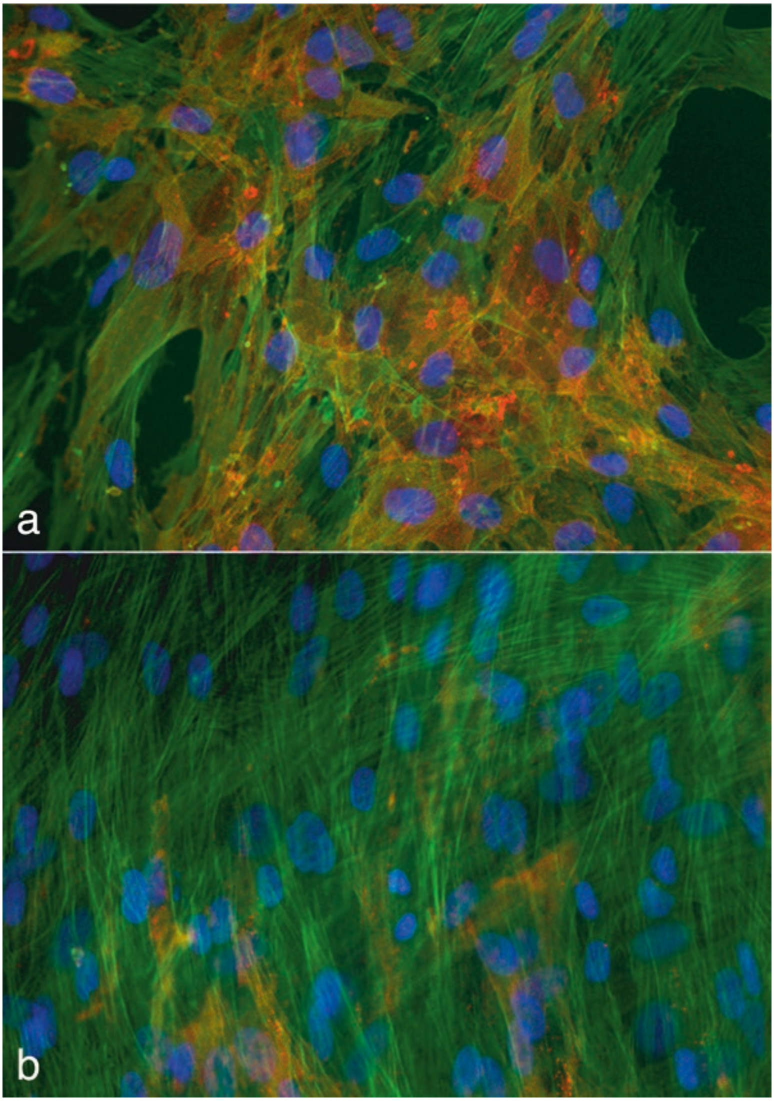


Fig. 9.2 Cell morphology and ALP expression in mesenchymal stem cells after 7 days of culture: (**a**) cells in contact with osteogenic medium, showing polygonal shape and expression of ALP; (**b**) elongated cells in the absence of osteogenic medium with few cells positive for ALP. Green labeling shows actin and blue stain labels cell nuclei (DAPI). Magnification of 400×, fluorescence microscopy

molecules. Type I collagen, fibronectin, and integrins are critical for osteoblast function and bone development (Cowles et al. 2000; Shekaran and Garcia 2011). Brunner et al. (2018) showed that the conditional deletion of β1 integrins in the osteo-precursor population severely impacts bone formation and homeostasis both

Table 9.1 Genes with shared modulation during osteoblast differentiation of bone marrow and umbilical cord vein mesenchymal stem cells (0–168 h in osteogenic medium). FDR ≤ 0.05, fold change ≥2.0

CloneID	GenBank acc	Gene name	Symbol	Cytoband	Function
22,074	AB209346	Thyroid hormone receptor, alpha-1	THRA	17q11.2	Transcription from RNA polymerase II promoter
23,536	NM_031461	Cysteine-rich secretory protein LCCL domain containing 1	CRISPLD1	8q21.11	Function unknown
23,586	NM_020708	Solute carrier family 12, (potassium-chloride transporter) member 5	SLC12A5	20q13.12	Potassium ion transport
24,032	BC037905	CASP2 and RIPK1 domain containing adaptor with death domain	CRADD	12q21.33-q23.1	Regulation of apoptosis
24,451	NM_021959	Protein phosphatase 1, regulatory (inhibitor) subunit 11	PPP1R11	6p21.3	Protein phosphatase inhibitor activity
24,630	NM_001184691	Nephronectin	NPNT	4q24	Cell differentiation
24,670	NM_101339	Purple acid phosphatase 3	PAP3	1	Acid phosphatase activity
25,220	AF131786	Clone 25,220 mRNA sequence	–	–	Function unknown
27,582	BX537526	CDNA FLJ11602 fis, clone HEMBA1003908	–	–	Function unknown
34,597	NM_001081550	THO complex 2	THOC2	Xq25-q26.3	MRNA-nucleus export
36,215	NM_004556	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon	NFKBIE	6p21.1	Cytoplasmic sequestering of transcription factor
53,042	AI220134	Transcribed locus	–	–	Function unknown
132,044	NM_002775	HtrA serine peptidase 1	HTRA1	10q26.3	Regulation of cell growth
132,702	NM_000918	Procollagen-proline, 2-oxoglutarate 4-dioxygenase beta subunit	P4HB	17q25	Electron transport
135,671	NM_002205	Integrin, alpha 5	ITGA5	12q11-q13	Cell-matrix adhesion
136,339	NM_198581	Zinc finger CCCH-type containing 6	ZC3H6	2q13	Function unknown
139,478	AK097984	Nicotinamide N-methyltransferase	NNMT	11q23.1	Transferase activity
140,268	BX396146	Hypothetical protein LOC728517	LOC728517	1p36.33	Function unknown
142,381	BQ278455	Eukaryotic translation initiation factor 1	EIF1	17q21.2	Regulation of translation

CloneID	GenBank acc	Gene name	Symbol	Cytoband	Function
142,568	NM_105885	Toxin receptor binding	THI2.1	1	
143,729	NM_198679	Rap guanine nucleotide exchange factor (GEF) 1	RAPGEF1	9q34.3	Transmembrane receptor protein tyrosine kinase signaling pathway
143,790	AL050366	O-linked n-acetylglucosamine transferase	OGT	Xq13	Signal transduction
154,117	BF570935	Lectin, galactoside-binding, soluble, 1 (galectin 1)	LGALS1	22q13.1	Apoptosis
259,888	NM_013355	Protein kinase N3	PKN3	9q34.11	Signal transduction
1,613,637	NM_006904	Protein kinase, DNA-activated, catalytic polypeptide	PRKDC	8q11	Protein modification

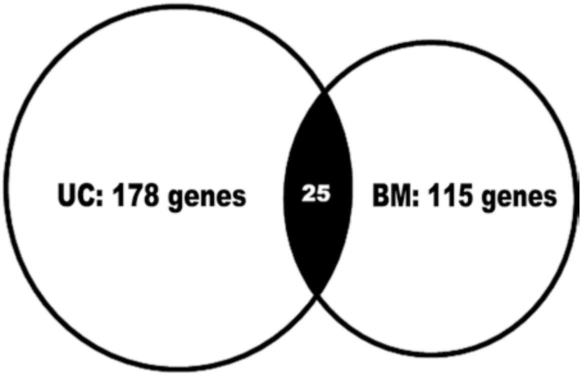


Fig. 9.3 Venn diagram showing the specific and the 25 sharing genes during normal differentiation of mesenchymal cells obtained from human bone marrow and umbilical cord vein. UC: umbilical cord vein cells; BM: bone marrow mesenchymal cells

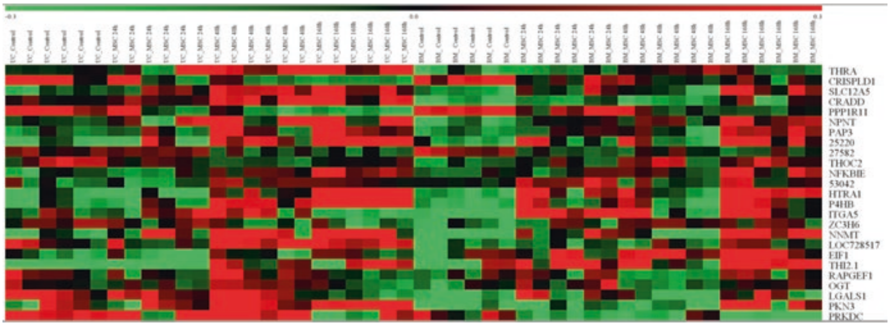


Fig. 9.4 Expression profiling of the 25 genes with shared modulation during osteoblast differentiation of bone marrow (BM) and umbilical cord vein (UC) mesenchymal stem cells [0–7 days (168 h) cultured in osteogenic medium]. FDR ≤0.05 and fold change ≥2.0

in vivo and in vitro. These authors observed that mutant mice display severe bone deficit characterized by bone fragility and reduced bone mass and that β1 integrins are required for proper BMP2 dependent signaling at the pre-osteoblastic stage, by positively modulating Smad1/5-dependent transcriptional activity at the nuclear level (Figs. 9.3 and 9.4).

The HtrA serine peptidase 1 (HTRA1) gene promotes the regulation of cell proliferation (SOURCE Database). It has been proposed that the HtrA1 protein regulates biological processes by modulating growth-factor systems other than IGF, such as the system mediated by the transforming growth factor beta 1 (TGFB1) family. Transforming growth factor beta (TGF-beta) is effective in regulating osteoblast proliferation, differentiation, bone matrix maturation, and cell-specific gene expression, as well as inhibiting the expression of markers characteristic of the osteoblast phenotype such as osteocalcin (Oka et al. 2004). Hadfield et al. (2008) suggested that HTRA1 may regulate matrix calcification via the inhibition of BMP-2 signaling, modulating osteoblast gene expression, and/or via the degradation of specific matrix proteins. Recent reports demonstrated that Htra1 is a positive regulator of osteogenic differentiation, showing that Htra1 is a direct downstream target of RUNX2 (Iyyanar et al. 2019).

Nephronectin (NPNT) gene codes an extracellular matrix protein highly expressed in long bones. Kahai et al. (2009) discovered that ectopic expression of nephronectin promotes osteoblastic differentiation, thus corroborating our results. Kuek et al. (2016) have shown that NPNT is expressed by osteoblasts and its expression is reduced in osteoporosis. Besides presenting a direct effect on endothelial cell activities and on the regulation of angiogenesis via p-38 and ERK pathways, NPNT is pointed out as a potentially important molecule in the communication between osteoblasts and endothelial cells by a paracrine mode of action.

The thyroid hormone receptor, alpha-1 (THRA) gene codes one of the several receptors of thyroid hormone, an established regulator of skeletal growth and maintenance both in clinical studies and in laboratory models (Lindsey et al. 2018). This gene is involved in the formation of bone or of a bony substance and the conversion of fibrous tissue or cartilage into bone or a bony substance (SOURCE Database).

Protein phosphatase 1 regulatory inhibitor subunit 11 (PPP1R11) was also found to be a shared gene. Considering that phosphatase activity is important for osteoblast differentiation, as in the case of ALPL that we determined in this study, and that the PPP1R11 protein is associated with inhibition of phosphatase activity, this may be evidence for a mechanism involving phosphatase enhancement/inhibition during osteoblast differentiation. Further, genes involved in kinase activity/protein phosphorylation such as Rap guanine nucleotide exchange factor (GEF) 1 (RAPEF1) and Protein kinase 3 (PKN3) also appeared, reinforcing the importance of phosphate metabolism in osteoblast differentiation.

Genes that control apoptosis such as lectin, galactoside-binding, soluble, 1 (LGALS1) and CASP2 and RIPK1 domain containing adaptor with death domain (CRADD) were also shared between the two sources of MSCs, providing evidence for controlled cell death during differentiation.

Genes involved in general processes such as control of transcription including nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon (NFKBIE) and control of ion transport as solute carrier family 12 (potassium/chloride transporter), member 5, (SLC12A5) were also shared.

Finally, we identified the participation of the protein kinase, DNA-activated, catalytic polypeptide (PRKDC) gene; in addition to its role in kinase activity and osteoblast differentiation as discussed above, this gene also plays roles in DNA and in the control of apoptosis, which are both processes that ultimately regulate cancer. These results suggest that regardless of the anatomical site from which stem cells were obtained, a shared set of genes is activated to trigger osteoblast differentiation.

Acknowledgments This study was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasília, Brasil), Project # 552227/2005-6 and # 305787/2017-9 to GAP.

References

Abdal Dayem A, Lee SB, Kim K et al (2019, 1922) Production of mesenchymal stem cells through stem cell reprogramming. Int J Mol Sci 20(8)

Abrams WR, Ma RI, Kucich U et al (1995) Molecular cloning of the microfibrillar protein MFAP3 and assignment of the gene to human chromosome 5q32-q33.2. Genomics 26:47–54

Aranda P, Agirre X, Ballestar E et al (2009) Epigenetic signatures associated with different levels of differentiation potential in human stem cells. PLoS One 13:e7809

Babb R, Chandrasekaran D, Neves VC, Sharpe PT (2017) Axin2-expressing cells differentiate into reparative odontoblasts via autocrine Wnt/β-catenin signaling in response to tooth damage. Sci Rep 7(1):3102

Billing AM, Hamidane HB, Dib SS, Cotton RJ et al (2016) Comprehensive transcriptomic and proteomic characterization of human mesenchymal stem cells reveals source specific cellular markers. Sci Rep 6:21507

Bombonato-Prado KF, Bellesini LS, Junta MM et al (2009) Microarray-based gene expression analysis of human osteoblasts in response to different biomaterials. J Biomed Mater Res A 88:401–408

Brunner M, Mandier N, Gautier T et al (2018) β1 integrins mediate the BMP2 dependent transcriptional control of osteoblast differentiation and osteogenesis. PLoS One 13(4):e0196021

Burns JS, Rasmussen PL, Larsen KH et al (2010) Parameters in three-dimensional osteopheroids of telomerized human mesenchymal (stromal) stem cells grown on osteoconductive scaffolds that predict in vivo bone-forming potential. Tissue Eng Part A 16:2331–2342

Campagnoli C, Roberts IA, Kumar S et al (2001) Identification of mesenchymal stem/progenitor cells in human first-trimester fetal blood, liver, and bone marrow. Blood 98:2396–2402

Carinci F, Piatelli A, Stabellini G et al (2004) Calcium sulfate: analysis of MG63 osteoblast-like cell response by means of a microarray technology. J Biomed Mater Res B Appl Biomater 71:260–267

Chen CT, Shih YRV, Kuo TK et al (2008) Coordinated changes of mitochondrial biogenesis and antioxidant enzymes during osteogenic differentiation of human mesenchymal stem cells. Stem Cells 26:960–968

Chen H, Liu X, Chen H et al (2014) Role of SIRT1 and AMPK in mesenchymal stem cells differentiation. Ageing Res Rev 13C:55–64

Cohen D, Melton D (2011) Turning straw into gold: directing cell fate for regenerative medicine. Nat Rev Genet 12:243–252

Cowles EA, Brailey LL, Gronowicz GA (2000) Integrin-mediated signaling regulates AP-1 transcription factors and proliferation in osteoblasts. J Biomed Mater Res 52:725–737

Cui Q, Xing J, Yu M et al (2019) Mmu-miR-185 depletion promotes osteogenic differentiation and suppresses bone loss in osteoporosis through the Bgn-mediated BMP/Smad pathway. Cell Death Dis 10(3):172

de Jong DS, Vaes BL, Dechering KJ et al (2004) Identification of novel regulators associated with early-phase osteoblast differentiation. J Bone Miner Res 19:947–958

de Souza LE, Malta TM, Kashima Haddad S, Covas DT (2016) Mesenchymal stem cells and pericytes: to what extent are they related? Stem Cells Dev 25(24):1843–1852

Deng ZL, Sharff KA, Song WX et al (2008) Regulation of osteogenic differentiation during skeletal development. Front Biosci 13:2001–2021

Edwin F, Patel T (2008) A novel role of Sprouty 2 in regulating cellular apoptosis. J Biol Chem 283:3181–3190. https://doi.org/10.1074/jbc.M706567200

Erices A, Conget P, Minguell JJ (2000) Mesenchymal progenitor cells in human umbilical cord blood. Br J Haematol 109:235–242

Fideles SOM, Ortiz AC, Assis AF et al (2019) Effect of cell source and osteoblast differentiation on gene expression profiles of mesenchymal stem cells derived from bone marrow or adipose tissue. J Cell Biochem 120:11842–11852

Guillot PV, De Bari C, Dell'accio F et al (2008) Comparative osteogenic transcription profiling of various fetal and adult mesenchymal stem cell sources. Differentiation 76:946–957

Hadfield KD, Rock CF, Inkson CA et al (2008) HtrA1 inhibits mineral deposition by osteoblasts: requirement for the protease and PDZ domains. J Biol Chem 283:5928–5938

Hofmann AD, Beyer M, Krause-Buchholz U et al (2012) OXPHOS supercomplexes as a hallmark of the mitochondrial phenotype of adipogenic differentiated human MSCs. PLoS One 7(4):e35160

Inkson CA, Ono M, Bi Y et al (2009) The potential functional interaction of biglycan and WISP-1 in controlling differentiation and proliferation of osteogenic cells. Cells Tissues Organs 189:153–157

Ishige I, Nagamura-Inoue T, Honda MJ et al (2009) Comparison of mesenchymal stem cells derived from arterial, venous, and Wharton's jelly explants of human umbilical cord. Int J Hematol 90:261–269

Iyyanar PPR, Thangaraj MP, Eames BF et al (2019) Htra1 is a novel transcriptional target of RUNX2 that promotes osteogenic differentiation. Cell Physiol Biochem 53(5):832–850

Jeong JA, Hong SH, Gang EJ et al (2005) Differential gene expression profiling of human umbilical cord blood-derived mesenchymal stem cells by DNA microarray. Stem Cells 23:584–593

Jongwattanapisan P, Terajima M, Miguez PA et al (2018) Identification of the effector domain of biglycan that facilitates BMP-2 osteogenic function. Sci Rep 8(1):7022

Kahai S, Lee SC, Lee DY et al (2009) MicroRNA miR-378 regulates nephronectin expression modulating osteoblast differentiation by targeting GalNT-7. Plos One 4: e7535

Kim CG, Lee JJ, Jung DY (2006) Profiling of differentially expressed genes in human stem cells by cDNA microarray. Mol Cells 21:343–355

Knight MN, Hankenson KD (2013) Mesenchymal stem cells in bone regeneration. Adv Wound Care (New Rochelle) 6:306–316

Kuek V, Yang Z, Chim SM et al (2016) NPNT is expressed by osteoblasts and mediates angiogenesis via the activation of extracellular signal-regulated kinase. Sci Rep 6:36210. https://doi.org/10.1038/srep36210. Erratum in: Sci Rep 21;6:37482

Kulterer B, Friedl G, Jandrositz A et al (2007) Gene expression profiling of human mesenchymal stem cells derived from bone marrow during expansion and osteoblast differentiation. BMC Genomics 8:70

Lindsey RC, Aghajanian P, Mohan S (2018) Thyroid hormone signaling in the development of the endochondral skeleton. Vitam Horm 106:351–381

Mareschi K, Biasin E, Piacibello W, Aglietta M, Madon E & Faioli F (2001). Isolation of human mesenchymal stem cells: bone marrow versus umbilical cord blood. Haematologica, 86:1099–1100

Ogura N, Kawada M, Chang W, Zhang Q , Lee S, Kondoh T, Abiko Y (2004) Differentiation of the human mesenchymal stem cells derived from bone marrow and enhancement of cell attachment by fibronectin. Journal of Oral Science, 46(4):207–213

Oka C, Tsujimoto R, Kajikawa M et al (2004) HtrA1 serine protease inhibits signaling mediated by Tgf beta family proteins. Development 131:1041–1053

Panepucci RA, Siufi JL, Silva WA et al (2004) Comparison of gene expression of umbilical cord vein and bone marrow-derived mesenchymal stem cells. Stem Cells 22:1263–1278

Pattappa G, Heywood HK, De Bruijn JD et al (2011) The metabolism of human mesenchymal stem cells during proliferation and differentiation. J Cell Physiol 226:2562–2570

Rajewsky N (2011) microRNAs and the operon paper (2011). J Mol Biol 409:70–75

Roson-Burgo B, Sanchez-Guijo F, Del Cañizo C et al (2016) Insights into the human mesenchymal stromal/stem cell identity through integrative transcriptomic profiling. BMC Genomics 17(1):944

Sacchetti B, Funari A, Remoli C, Giannicola G, Kogler G, Liedtke S, Cossu G, Serafini M, Sampaolesi M, Tagliafico E, Tenedini E, Saggio I, Robey PG, Riminucci M, Bianco P (2016) No identical "mesenchymal stem cells" at different times and sites: human committed progenitors of distinct origin and differentiation potential are incorporated as adventitial cells in microvessels. Stem Cell Rep 14;6(6):897–913

Sarugaser R, Lickorish D, Baksh D et al (2005) Human umbilical cord perivascular (HUCPV) cells: a source of mesenchymal progenitors. Stem Cells 23:220–229

Schilling T, Küffner R, Klein-Hitpass L et al (2008) Microarray analyses of transdifferentiated mesenchymal stem cells. J Cell Biochem 103:413–433

Secco M, Moreira YB, Zucconi E et al (2009) Gene expression profile of mesenchymal stem cells from paired umbilical cord units: cord is different from blood. Stem Cell Rev 5:387–401

Semeghini MS, de Azevedo FG, Fernandes RR et al (2018) Menopause transition promotes distinct modulation of mRNAs and miRNAs expression in calvaria and bone marrow osteoblastic cells. Cell Biol Int 42(1):12–24

Shekaran A, Garcia AJ (2011) Extracellular matrix-mimetic adhesive biomaterials for bone repair. J Biomed Mater Res A 96:261–272

Shen YS, Chen XJ, Wuri SN et al (2020) Polydatin improves osteogenic differentiation of human bone mesenchymal stem cells by stimulating TAZ expression via BMP2-Wnt/β-catenin signaling pathway. Stem Cell Res Ther 11(1):204

Shi S, Robey PG, Gronthos S et al (2001) Comparison of human dental pulp and bone marrow stromal stem cells by cDNA microarray analysis. Bone 29:532–539

Stadtfeld M, Hochedlinger K (2010) Induced pluripotency: history, mechanisms, and applications. Genes Dev 24:2239–2263

Tanabe S (2014) Role of mesenchymal stem cells in cell life and their signaling. World J Stem Cells 26:24–32

Tormos KV, Anso E, Hamanaka RB et al (2011) Mitochondrial complex III ROS regulate adipocyte differentiation. Cell Metab 14:537–544

Ulloa-Montoya F, Kidder BL, Pauwelyn KA et al (2007) Comparative transcriptome analysis of embryonic and adult stem cells with extended and limited differentiation capacity. Genome Biol 8:R163

Vesela B, Svandova E, Hovorakova M et al (2019) Specification of Sprouty2 functions in osteogenesis in in vivo context. Organogenesis 15(4):111–119

Wagner W, Wein F, Seckinger A et al (2005) Comparative characteristics of mesenchymal stem cells from human bone marrow, adipose tissue, and umbilical cord blood. Exper Hematol 33:1402–1416

Wang HS, Hung SC, Peng ST et al (2004) Mesenchymal stem cells in the Wharton's jelly of the human umbilical cord. Stem Cells 22:1330–1337

Welsh IC, Hagge-Greenberg A, O'Brien TP et al (2007) A dosage-dependent role for Spry2 in growth and patterning during palate development. Mech Dev 124:746–761

Wexler SA, Donladson C, Denning-Kendall P et al (2003) Adult bone marrow is a rich source of human mesenchymal 'stem' cells but umbilical cord and mobilized adult blood are not. Br J Haematol 121:368–374

Xu L, Liu Y, Sun Y et al (2017) Tissue source determines the differentiation potentials of mesenchymal stem cells: a comparative study of human mesenchymal stem cells from bone marrow and adipose tissue. Stem Cell Res Ther 8(1):275

Yaghoubi Y, Movassaghpour A, Zamani M et al (2019) Human umbilical cord mesenchymal stem cells derived-exosomes in diseases treatment. Life Sci 233:116733

Yianni V, Sharpe PT (2019) Perivascular-derived mesenchymal stem cells. J Dent Res 98(10):1066–1072

Zhao J, Faure L, Adameyko I, Sharpe PT (2021) Stem cell contributions to cementoblast differentiation in healthy periodontal ligament and periodontitis. Stem Cells 39(1):92–102. https://doi.org/10.1002/stem.3288.

Zhou LL, Liu W, Wu YM et al (2020) Oral mesenchymal stem/progenitor cells: the immunomodulatory masters. Stem Cells Int 2020:1327405

Chapter 10
Transcriptomics to Dissect the Immune System

Hideyuki Yoshida, Mitsuru Matsumoto, and Minoru Matsumoto

Immunology, which is the study for the immune system, started in the late nineteenth century beginning with two significant discoveries. One was the phagocytosis by macrophages which plays a critical host-defense mechanism against invading pathogens found by Elie Metchnikoff (1845–1916). The other one was an antibody which can neutralize microbial toxins discovered by Emil von Behring (1854–1917) and Paul Ehrlich (1854–1915) (Kaufmann 2017). Since then, immunology has been a field of intensive biomedical research and contributed to society by providing pivotal knowledge on both basic science and clinical applications along with its development. While the classical and authentic function of the immune system is to protect our bodies from diverse pathogenic microorganisms, including bacterias, viruses, and parasites, recent immunological studies revealed different parts of the immune system in eliminating cancer cells and regulating physiologic processes in diverse tissues such as the nervous system function, metabolic state, thermogenesis, and tissue repair (Chaplin 2010; Rankin and Artis 2018; Rouse and Sehrawat 2010). Now, we recognize the immune system is a multifunctional biological system and vital for our health and survival.

H. Yoshida (✉)
YCI Laboratory for Immunological Transcriptomics, RIKEN Center for Integrative Medical Science, Yokohama, Japan
e-mail: hideyuki.yoshida@riken.jp

M. Matsumoto
Division of Molecular Immunology, Institute for Enzyme Research, Tokushima University, Tokushima, Japan

M. Matsumoto
Division of Molecular Immunology, Institute for Enzyme Research, Tokushima University, Tokushima, Japan

Department of Molecular Pathology, Tokushima University Graduate School of Biomedical Sciences, Tokushima, Japan

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_10

The immune system is composed of different types of immune cells, which do not form a single organ like the brain and heart but are spread throughout the body to achieve rapid responses to invading pathogens. As transcriptional regulation plays a crucial role in shaping these immune cells of diverse differentiation and activation status, various immune cells were examined at the transcriptional level. These profiling analyses effectively yielded relevant insights of immune cells regarding respective regulatory mechanisms and crucial factors involved in cell differentiation and activations (Amit et al. 2011; Lara-Astiaso et al. 2014; Mostafavi et al. 2016; Smale and Fisher 2002; Uhlen et al. 2019). Furthermore, recent single-cell transcriptome analyses by single-cell RNA sequencing (scRNA-seq) provide unprecedented high-resolution insights of immune cells which cannot be captured by studies in bulk and are expected to promote our understanding of the nature of immune cells in both physiological and pathological contexts (Proserpio and Mahata 2016; Roy 2019; Seumois and Vijayanand 2019; Stubbington et al. 2017; Xie et al. 2021). We introduce general features of the immune system and discuss the transcriptome analysis applied to explore the immune system.

10.1 The Immune System and Immune Cells

The immune system not only protects us from diverse infections, including bacterias, viruses, and parasites, but also eliminates cancer cells and healing wounds. The efficiency of the immune activity relies on the orchestrated functions of a set of different types of immune cells, which are responsible for the diverse steps of the process. In the case of infections, for example, these include pathogen recognition, the cascade to recruit and activate effector cells, and the final clearance by other immune cells. Identifications of different types of cells involved in the immune process have been a keen target of immunological research for decades, and accordingly, types of immune cells have been expanded, which contributed to the dissection of the immune functions. It was started from the discovery of white blood cells in 1843 by Gabriel Andral (1797–1876) and William Addison (1802–1881). Then, different types of immune cells have been progressively identified along with the development of technologies such as flow cytometry in the 1960s and monoclonal antibodies in the 1970s, which were collectively employed to specify CD4$^+$ T cells and CD8$^+$ T cells, for instance (Hajdu 2003; Jayasinghe 2020; Packer 2021). The major populations of immune cells include granulocytes and macrophages with innate ability to phagocytose bacteria, antibody-producing B cells which were discovered before the 1990s, and more than 80 immune cell populations are recognized to date (Fig. 10.1) (Ackerman 1964; Hayakawa et al. 1983; Maecker et al. 2012; Stein et al. 1992).

While different immune cells possess distinctive functions, essentially all immune cells develop from a hematopoietic stem cell in the bone marrow and share the same genome except for rearranged genes (i.e., T cell receptor and immunoglobulin). Through the differentiation pathways that can be parsed up to as

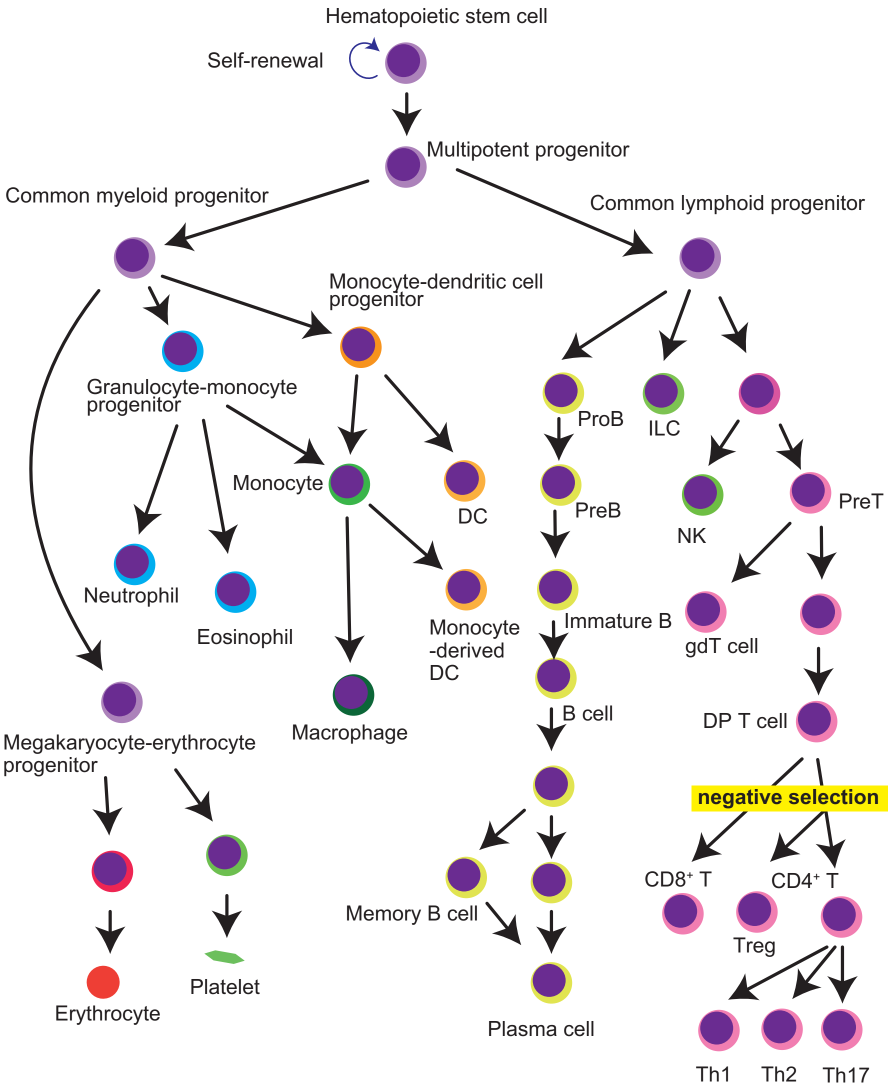


Fig. 10.1 Overview of immune cell populations
Arrows indicate schematic representation of the standard model for hematopoietic stem cell differentiation. Self-reactive T cells are eliminated by negative selection after the CD4/CD8 double-positive (DP) stage in the thymic medulla with the help of mTECs. *ILC* **innate lymphoid cell,** *NK* **natural killer cell,** *gdT* **gamma delta (γδ) T cell,** *Treg* **regulatory T cell,** *DC* **dendritic cell**

many as ten successive steps, immune cells acquire their divergent capabilities, which are established by correspondent transcriptional landscapes (Hardy and Hayakawa 2001; Rothenberg 2014). As such, transcriptional regulations are the most fundamental mechanisms controlling immune cells and the immune system. Thus, it is not surprising that the recent development of single-cell transcriptomics is promising to confirm existing populations and unveil new populations efficiently in an unbiased manner.

10.2 Transcriptome Analysis of Different Subsets in Bulk

While more than 80 immune cell subsets are recognized throughout our body, many subsets residing in lymph nodes, tissues, and organs, immunocytes in peripheral blood are the most feasible cells to be examined for research and clinical diagnostics (Chou and Li 2018; Maecker et al. 2012; Novershtern et al. 2011). A few large transcriptomic studies have been done on different immune cell populations in blood. For example, 13 immune cell types in peripheral blood were examined by Schmiedel et al. (2018) 29 immune cell types by Monaco et al. (2019), and 18 immune cell types by Uhlen et al. (2019). In these studies, they isolated immune cells in blood including such as monocytes, natural killer (NK) cells, neutrophils, basophils, B cells, CD4$^+$ and CD8$^+$ T cells, as well as dendritic cells (DCs) employing known markers and fluorescence-activated cell sorters (FACS), and profiled whole transcriptomes by RNA sequencing (RNA-seq) or microarrays. These studies have revealed the distinctive global expression profiles of various immune cells where granulocyte cell types (neutrophils, basophils, and eosinophils) are discrete from others, all lymphocytes make a cluster, including T cells, NK cells, and B cells. In contrast, monocytes are closely related to DCs. According to the study by Uhlen et al., among ~16,000 genes detected in 18 immune cell types, ~10,000 were detected in single-cell types, which were almost comparable to genes detected in cell lines (~9,500 genes per cell line). Of these, 5,934 genes were seen across all immune cells and 1,713 genes in a single-cell type: 9,939 genes showed low specificity for cell types. The sets of differentially expressed and co-expressed genes were served to deduce the functional modules of genes with the aid of bioinformatics such as enrichment analysis using the gene ontology (GO). Furthermore, these transcriptome atlases in each cell type are valuable to promote the understanding of primary immunodeficiency diseases (PID). PID are a large group of over 400 different diseases caused by quantitative and functional changes in the various mechanisms involved in immune response and associated with complications including infections, autoimmune disorders, immune dysregulation with lymphoproliferation, inflammatory disorders, lymphomas, and other types of cancers (Amaya-Uribe et al. 2019; Sánchez-Ramón et al. 2019). While PID are caused by genetic disorders and 354 diseases were listed as consequences of monogenic defects in genes associated with the immune system involving 224 identified genes, the mechanism of disease is often incompletely understood (Uhlen et al. 2019) (https://www.omim.org). Uhlen et al. hypothesized that an analysis of cellular expression of identified genes could help generate a better mechanistic investigation and analyzed 224 PID genes across their 18 immune cell populations.

They divided these PID genes into seven clusters according to the shared expression pattern among cell populations and found some PID genes are expressed explicitly in restricted populations. These included the CEBPE gene in which mutations can cause specific granule deficiency 1 (SG1) highly expressed in eosinophils. Although SG1 has been considered a neutrophil-granule deficiency associated with recurrent pyogenic infections, CEBPE's expression in eosinophils suggested that eosinophil deficiency might also be involved in SG1.

It is also worth noting that the variable mRNA abundance in different immune cells was carefully examined in the study by Monaco et al., and they developed an enhanced method for normalization. The normalization for mRNA abundance can become essential for differential expression analyses. For example, if the analysis is done with two cell types of essentially different total mRNA amounts per cell (e.g., same 10 gene X mRNA molecules are expressed, but cell type A is expressing 100 mRNA molecules in total, and cell type B is expressing 1,000 mRNA molecules in total), this can lead to the misleading that the gene X is downregulated in cell type B. Indeed, existing normalization methods for transcriptome profiling such as the UQ, TMM, and RLE cannot correctly identify transcriptomes in which the overall transcriptional activity is suppressed or enhanced (Anders and Huber 2010; Bullard et al. 2010; Robinson and Oshlack 2010). Normalizing mRNA abundance also becomes relevant to analyzing the transcriptomes from cells of heterogeneous populations such as peripheral blood mononuclear cells (PBMCs) by employing a deconvolution method. Deconvolution processing computationally estimates the proportions of distinctive cell types in a heterogeneous sample utilizing the normalized abundance of mRNA in each cell type as references It is an effective solution to determine the composition of each immune cell type in PBMCs (Abbas et al. 2009; Shen-Orr and Gaujoux 2013). Considering that the proportion of immune cell subsets in PBMCs can be dynamically affected by the disease, age, or interventions (e.g., vaccines and drugs), the composition of immune cell populations needs to be carefully evaluated. Otherwise, it is not always possible to accurately determine which immune cell types are responsible for any given transcriptomic changes in PBMCs. The transcriptome profiling can contribute to the results that are inconclusive or difficult to interpret. Hence, the appropriate normalization method is crucial for differential expression analyses and deconvolution approaches. Monaco et al. developed an advanced and robust normalization method that can be applied for future transcriptome analyses of PBMCs by taking advantage of the breadth and granularity of the datasets from 29 isolated immune cell types.

In summary, transcriptome analyses of isolated immune cells from peripheral blood elucidate individual immune cell population' divergent gene expression patterns, which promote our understanding of diseases related to the immune system. The transcriptome analyses of isolated immune cells are also critical as the resource for analyzing transcriptomes obtained from whole peripheral blood. Furthermore, large transcriptomic studies of isolated immune cells provide opportunities to develop and validate analysis pipelines which would be impractical from heterogeneous samples.

Another point to mention here is that the immune cells have been exploited to investigate the regulatory mechanisms of gene expressions. The immune system serves as an excellent model to explore the gene regulations along changing cell states, as discrete cell populations can be readily purified by well-established markers along differentiation and activation pathways that have been carefully characterized by persuasive studies. We took advantage of the breadth and granularity of immune cells to study the dynamic epigenetic landscapes associated with the target gene expression (Yoshida et al. 2019). The study provided a deep insight to understand immunological differentiation and function and the broad relevance of gene regulatory elements on the genome, such as a profound dichotomy within mammalian gene regulation by enhancers and promoters.

10.3 Transcriptomes Analyses Employing Whole Blood and PBMC

Blood is an invaluable source to examine our health not only because of the easy accessibility and minimal invasiveness during sampling but also because of the breadth of information it can provide (Sohn 2017). Transcriptomes in PBMCs have also been investigated intensively for scientific research (Corkum et al. 2015; Mello et al. 2012), as well as in medical contexts such as ischemic stroke (Baird et al. 2015), ulcerative colitis (Miao et al. 2013), epilepsy (Karsten et al. 2011), and sepsis (Davenport et al. 2016) to characterize diseases, and epidemiological contexts including aging (Peters et al. 2015), obesity (Homuth et al. 2015), and lifestyle factors such as smoking, drinking, and nutrition (Burton et al. 2018; Dumeaux et al. 2010).

Since PBMCs can include variable naïve and activated immune cells recirculating throughout the body, PBMCs transcriptome analyses are expected to promote the characterization of the whole immune system. However, due to the heterogeneity and dynamics of the components of immune cell types in PMBCs, cell population-level resolution is not successfully achieved so far even with the cutting-edge approaches such as deconvolutions mentioned and thus straight immunological interpretations (e.g., a specific immune cell population is enlarged in donor A than donor B, or a set of genes are more activated in immune cell population X in donor A than B) are readily possible. Accordingly, different approaches employing systems biology are preferentially applied for analyzing transcriptomes from PBMCs (Chaussabel 2015).

Systems biology is an approach in the biomedical research field to understand the larger picture hidden in the biological system by putting pieces of information from the system together. A hypothesis being constructed based on all observed parameters associated with a given biological system, systems biology is compatible with high throughput technologies called "omics" such as genomics, transcriptomics, proteomics, and metabolomics by which a biology system is comprehensively profiled (Aizat et al. 2018; Veenstra 2021). In omics, the parameters are not chosen in advance like in more traditional assays, and these approaches are inherently unbiased. Importantly, as the potency of systems biology intrinsically relies on the variability of observed parameters, the size and heterogeneity of a dataset are crucial for the analyses employing systems biology, and thus more informative results can be expected from the larger dataset (Koumakis 2020; McCue and McCoy 2017; Qin et al. 2015). Schmidt et al. reported the analysis of blood transcriptomes of 3,388 adult individuals (mean age = 58 years), together with phenotypic attributes including disease history, medication status, lifestyle factors, and body mass index (BMI) (Schmidt et al. 2020). Although there were preceding studies analyzing blood transcriptomics, studies were composed of relatively smaller sample sizes related to specific diseases, which restricted the analytical power due to the limited variability in the transcriptomic states and health conditions. Schmidt et al. demonstrated the diversity of blood transcriptomes with modules of co-expressed genes linking to different biological functions. They visualized the molecular heterogeneity of transcriptomes combining with different phenotypic statuses by employing state-of-the-art machine learning methods. The results include two major transcriptomic types, one relating to inflammation enhanced in male, elderly, and overweighted people, and the other one to activated immune responses in female, younger, and ordinary weighted people. They also found that transcriptome signatures are associated with immune response and the increase of inflammatory processes are shared among multiple diseases, aging, and obesity, indicating common underlying mechanisms.

Together, transcriptome analyses employing blood or PBMC is not straightforward to elucidate biological processes at the cell population level and characterize specific immune processes. However, they can provide an unprecedented opportunity to evaluate various diseases and lifestyle factors. They will be applicable for medical diagnostics and molecular and epidemiological research, which will contribute to the promotion of the personalized medicine.

10.4 Transcriptome Analyses: From Bulk to Single Cells

As mentioned earlier, transcriptome analyses using isolated immune cells as well as blood cells are beneficial to promote our understanding of the immune system by shedding light on disease pathogenesis and global immunity. However, as these are averaged profiles of immune cells and the transcriptomes of minor cells are masked by other major cells, it is not feasible to detect its relevance if rare subsets of cells are responsible for an immune phenotype. The heterogeneity is evident in blood cells and PMBC. Still, FACS-isolated cells according to their markers can also be heterogeneous because immune cell types are too heterogeneous to be entirely separated by known markers. Furthermore, immune cells can be activated by various stimuli such as pathogens and secreted proteins from other cell types (i.e., cytokines) temporarily in an unsynchronized manner. The heterogeneity should also be considered when rare subsets of cells (e.g., antigen-specific T cells or B cells) drive the immune responses by temporal activation (Chattopadhyay et al. 2014; Mostafavi et al. 2016). Hence, single-cell analysis is most anticipated when seeking rare distinctive subsets of cells relating to biological outcomes, for example, when rare cells are essential for conferring protection or inducing pathologic status.

10.5 Recent Development of Single-Cell Transcriptomics

Before the transcriptome analyses from single cells became possible, cDNA synthesis and amplification from a single cell were first succeeded by Iscove in 1990 and Coleman in 1992 (Brady et al. 1990; Eberwine et al. 1992). The cDNA was analyzed using DNA microarrays in the early 2000s and subsequently combined with next-generation sequencing (NGS) technology for single-cell RNA sequencing (scRNA-seq) around 2010 (Bengtsson et al. 2005; Islam et al. 2011; Klein et al. 2002; Kurimoto et al. 2006; Tang et al. 2009). There have been various scRNA-seq methods developed ranging from relatively lower throughput but more detailed full-length transcriptomic data from individual cells to higher throughput with focused coverage on the 3′ terminal of the transcript (Jaitin et al. 2014; Klein et al. 2015; Macosko et al. 2015; Picelli et al. 2013). Currently, scRNA-seq employing a commercial kit from 10× Genomics (Pleasanton, CA) is presumably the most popular. They allow us to profile up to 10,000 cells at a time, and have been used in more than 1,000 publications (Daniloski et al. 2021; Stewart et al. 2020).

10.6 Applying scRNA-seq for Immune Cells

The heterogeneity in immune cells mirrors the unusual flexibility of the immune system and is essential to protect our bodies efficiently from diverse pathogens. It was recognized in the 1970s and successively confirmed along with the identifications of the cluster of differentiation (CD) antigens using monoclonal antibodies (Engel et al. 2015; Talal 1973). For example, a type of T cells marked by CD4 glycoprotein molecule on the cell surface was identified around 1980. Then subtypes including Th1, Th2, Th17, and the regulatory T cells (Tregs) were identified later (Engleman et al. 1981; Harrington et al. 2005; Mosmann et al. 1986; Park et al. 2005; Sakaguchi et al. 1995). However, given that these distinctions between subtypes are defined by the expression of a few specific markers, these classifications might be a very simplified categorization. Indeed, Teichmann and colleagues demonstrated a subpopulation in Th2 cells which produces the steroid pregnenolone by employing the scRNA-seq approach (Mahata et al. 2014). Importantly, as the comprehensive transcriptome analysis was accomplished by scRNA-seq, they could identify co-regulated genes in the subpopulation, which facilitated the characterization of the cells. Shalek et al. also reported the transcriptomic heterogeneity within bone-marrow-derived dendritic cells (BMDCs) which were seemingly homogenous using scRNA-seq (Shalek et al. 2013). They found hundreds of key immune genes, including genes very highly expressed at the population level, are bimodally expressed across cells. While these pioneering researches employed mouse cells, scRNA-seq approaches were also effectively applied to human cells later.

Karamitros et al. employed scRNA-seq to investigate the transcriptomic differences between progenitor populations in human cord blood (i.e., lymphoid-primed multipotential progenitors: LMPPs, granulocyte-macrophage progenitors: GMPs and multi-lymphoid progenitors: MLPs which were FACS-isolated according to known markers) (Karamitros et al. 2018). They revealed these progenitors were transcriptionally distinct and heterogeneous at the single-cell level, with cells from different progenitor populations showing a transcriptional continuum. Combining with the results from functional assays, they argued a continuum of progenitors executed lymphoid and myeloid differentiation, rather than progenitors downstream of stem cells are uni-lineage. Considering that functional assays can only demonstrate the potential rather than actual cell fate in vivo, and a failure to display functional potential might reflect the assay's problem, transcriptome analysis adequately contributed to declaring progenitor's fate in vivo. Recently, Xie et al. profiled 7,551 human blood cells isolated from 21 healthy donors (Xie et al. 2021). They isolated 32 immunophenotypic cell types by FACS and measured transcriptomes in single cells by scRNA-seq. These cells include hematopoietic stem cells, progenitors, and mature immune cells, representing the whole-blood system. The transcriptomic profiles from these 7,551 cells constitute a comprehensive atlas for hematopoietic cells at single-cell resolution. Besides they identified putative long non-coding RNAs (lncRNAs) and transcription factors regulating the differentiation of immune cells, the atlas is also valuable as a resource. It will be utilized by the community to understand the transcriptomic regulations underlying hematopoiesis and immune cell differentiation.

10.7 scRNA-seq Analysis in Diseases

Measuring the transcriptomes at single-cell resolution by scRNA-seq is innovating our understanding of immune cells in a physiological setting, as mentioned above. In addition, this approach has afforded new options to study the immune response in pathological conditions. What types of cells are responsible for the dysregulated immune response in diseases? By employing comprehensive transcriptome profiling at single-cell resolution, it is possible to examine whether new pathogenic cell subsets developed in disease and the expansion (or contraction) of physiological cell subsets are accompanied. For example, Golumbeanu et al. employed the scRNA-seq approach for dissecting HIV-infected primary CD4$^+$ T cells (Golumbeanu et al. 2018). HIV can persist in latently infected cells despite the effective treatments, which hampers HIV eradication. Hence strategies so-called "shock and kill" have been developed aiming at reactivating HIV production from the latent cells, so as these cells will die due to virus-mediated cytotoxicity and be killed by cytotoxic CD8$^+$ T cells. However, reactivations of HIV expression are limited to a fraction of latent cells, and the heterogeneity of latently infected cells was suggested. Golumbeanu et al. identified two major cell subpopulations characterized by a set of 134 differentially expressed genes (DEGs) by employing scRNA-seq. Gene ontology analysis revealed enrichment of viral processes, translational regulation, RNA and protein metabolism as well as cell activation genes among these DEGs, which indicates different HIV reactivation potentials for each cluster. They argue that these DEGs are valuable to facilitate the identification of successful reactivations and to identify potential biomarkers of inducible cells.

The composition of the tumor microenvironment (TME) is known to affect the prognosis of cancer patients. For example, higher infiltrates of cytotoxic and memory CD8^{+} T cells, Th1 CD4^{+} cells, and NK cells are usually associated with better outcomes, whereas Th2 and Th17 CD4^{+} cells and Tregs with poor prognosis in several cancers (Fridman et al. 2012). Indeed, while immunotherapies for lung cancer can significantly improve the prognosis for patients, their efficacy varies and depends on in part the number and properties of tumor-infiltrating T cells. Guo et al. investigated the heterogeneity within the tumor-infiltrating T cell by scRNA-seq (Guo et al. 2018). They performed scRNA-seq for 12,346 T cells from 14 untreated non-small-cell lung cancer (NSCLC) patients to comprehensively understand the infiltrating T cells regarding composition, lineage and functional status, and demonstrated the heterogeneity within exhausted CD8^{+} T cells and Tregs. T-cell exhaustion was originally identified in mice during chronic infection and was later observed in cancer patients (Jiang et al. 2015; Pauken et al. 2016). Exhausted T cells in TME are hyporesponsive states expressing increased inhibitory receptors and decreased effector cytokines, which provoke the failure of cancer elimination. Reinvigorating T-cell exhaustion by such as anti-CTLA-4 (ipilimumab) and anti-PD-1 (nivolumab and pembrolizumab) represents a promising strategy to treat cancer. Since scRNA-seq facilitates trajectory inference or so-called pseudo-time ordering which estimates the cellular identity along with a consecutive differentiation without prior knowledge, they could analyze CD8^{+} T cells undergoing exhaustion in TME and anticipate two clusters of cells preceding exhaustion, including their transcriptome signatures (Saelens et al. 2019). They employed the transcriptome datasets from TCGA LUAD (The Cancer Genome Atlas Lung Adenocarcinoma) and demonstrated that a high ratio of pre-exhausted to exhausted T cells was associated with a better prognosis. Furthermore, they identified heterogeneity within Tregs in TME, marked by the bimodal expression pattern of *TNFRSF9* which is a known activation marker for Tregs. They found a set of 260 genes, including *REL* and *LAYN* which are associated with immunosuppressive functions, are highly expressed in *TNFRSF9*$^{+}$ Tregs compared to *TNFRSF9*$^{-}$ Tregs. Importantly, survival analysis employing the TCGA LUAD dataset indicated that higher expressions of these 260 genes were predictive of a worse prognosis. These results represent the efficacy of an approach using scRNA-seq to reveal the heterogeneity in immune cell populations and identify potential clinical biomarkers.

Compared with bulk RNA-seq, scRNA-seq detects the transcriptome nuance in single cells that contribute to revealing the heterogeneity in a seemingly single population. With state-of-the-art machine learning and big data analytics, scRNA-seq has been becoming valuable to identify unknown subpopulations and their transcriptome signatures that affect the biological process and disease diagnosis. However, it is worth noting that scRNA-seq also has limitations compared with bulk RNA-seq, which include relatively low sensitivity, the bias of the transcriptome coverage, and overall cost. Hence, we anticipate that bulk RNA-seq will not lose its value.

10.8 The Paradigm of Self vs. Non-self from a Transcriptomic Viewpoint

Heterogeneity in the immune cells includes diversity at the DNA level besides the RNA and protein levels which establish the heterogeneity on the population level as discussed. At the DNA level, the number of T-cell receptors (TCRs) and the B-cell receptors (BCR) are estimated to be in the order of 10^7 whereas the human genome contains roughly 30,000 genes (Fugmann et al. 2000; Nikolich-Zugich et al. 2004). These are produced by somatic DNA recombination called V(D)J recombination in developing lymphocytes during the early stages of T and B cell differentiation. The exceptional divergency endows the immune system with potent effector mechanisms to destroy and eliminate a broad range of pathogenic microorganisms. As the recombination is nearly random, which appropriate to achieve the reactivity to targets of essentially unlimited diversities, it also causes the possibility of self-reactivity at the same time. Therefore, it is critical for the immune system to have mechanisms discriminating self from non-self to avoid destroying the host's own tissues. The capability of the immune system to avoid damaging the host's tissues is known as self-tolerance. As the failure of self-tolerance is associated with various autoimmune diseases, this mechanism has been broadly studied in immunology (Besnard et al. 2021; Klein et al. 2014; Sakaguchi et al. 2020).

One of the pivotal roles of T cells is to recognize and kill host cells infected by microbes which otherwise serve as factories for producing replicated microbes. This is managed by a mechanism where infected cells present a molecular complex of microbe antigens and Major Histocompatibility Complex (MHC) class I molecules on the cell surface, which are recognized and killed by T cells with a compatible TCR. As MHC molecules also present normal self-peptides on the cell surface, it is crucial for T cells to maintain self-tolerance. Negative selection of self-reactive T cells is an important process in the thymus where developing T cells of self-reactivity are eliminated if their TCRs react to self-peptides on MHC molecules. Intriguingly, essentially all protein-coding genes are expressed in sets of cells in the thymus. Negative selection functions effectively and comprehensively in the thymus where thymic epithelial cells (TECs) play a pivotal role. In the final section, we discuss how the establishment of self-tolerance in the thymus has been studied using transcriptomic data obtained by novel technologies.

10.8.1 Thymic Epithelial Cells (TECs) in the Thymic Stroma

The thymus is a highly specialized organ for the establishment of self-tolerance, which is characterized by the "education" of immature T cells. Thymus' key function is to provide diverse competent T cells that can recognize and eliminate foreign antigens, while they are tolerant to self-components. This complicated process is mainly orchestrated by TECs that form reticular structures in the thymus. TECs are divided into two major subsets by their localization, molecular characteristics, and functions: cortical TECs (cTECs) and medullary TECs (mTECs). Specifically, cTECs are responsible for T-cell lineage commitment and positive selection, while mTECs contribute to the negative selection of self-reactive T cells and/or their cell-fate diversion into Treg lineages (Kyewski and Klein 2006; Matsumoto et al. 2019). These incomparable roles in mTECs are achieved by the expression and presentation of diverse self-antigens complexed with MHC molecule on the surface of mTECs. Notably, to effectively screen for considerable self-reactive thymocyte clones, mTECs are equipped with a unique capacity to express almost 90% of the coding genome, including thousands of tissue-restricted antigens (TRAs) (Kadouri et al. 2020). As expected, the impairment of this "central tolerance" machinery can result in various autoimmune diseases. However, most autoimmune diseases are multifactorial, making it difficult to elucidate their pathogenesis. In this regard, autoimmune regulator (AIRE) and forkhead box P3 (FOXP3), both of which work as transcription factors, are very characteristic genes that cause severe autoimmunity by a single gene mutation. Considering its intimacy with TECs, we focus on and review Aire, an intriguing transcriptional regulator.

10.8.2 Aire in mTEC

The human *AIRE* gene was first cloned as the causative gene for autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) (Finnish-German APECED Consortium 1997; Nagamine et al. 1997). APECED shows autosomal recessive inheritance and patients have been preferentially reported in certain populations such as Finns, Norwegians, Sardinians, and Iranian Jews (Myhre et al. 2001). The human *AIRE* gene is composed of 14 exons and is located in the region q22.3 of chromosome 21, encoding a 545 amino-acid protein with a molecular weight of 57.5 kDa (Pitkanen et al. 2000). Importantly, Aire is almost exclusively expressed in mTECs in the thymus.

APECED patients' symptoms are characterized by a variable combination of (i) failure of the endocrine organs, (ii) chronic mucocutaneous candidiasis, and (iii) dystrophy of the ectoderm-derived tissues (Ahonen et al. 1990). The "hypoparathyroidism," "adrenal insufficiency (Addison disease)," and "chronic mucocutaneous candidiasis" are regarded as the triad of APECED. Notably, APECED patients have high levels of serum autoantibodies reacting specifically with components in the affected organs, like antibodies against steroidogenic enzymes of the P450 superfamily (e.g., P450c21 and P450c17) in the adrenal cortex (Peterson and Peltonen 2005). Furthermore, unique neutralizing autoantibodies to type I IFN and Th17-related cytokines are frequently detected in patients and these antibodies had been considered to be responsible for the development of chronic mucocutaneous candidiasis (Kisand et al. 2010; Puel et al. 2010). However, this long-standing hypothesis was recently challenged by another group, arguing that aberrantly enhanced type 1 immunity in the patients promotes candida infection susceptibility (Break et al. 2021).

Following the identification of the human *AIRE* gene, Aire-knockout (Aire-KO) mice (B6 genetic background) were generated to elucidate the mechanisms underlying the Aire deficiency and breakdown of self-tolerance (Anderson et al. 2002). Although the Aire-KO mice showed a rather milder phenotype than APECED patients, they developed lymphocytic infiltrates in several organs with the production of several autoantibodies. Remarkably, Aire-deficient mTECs showed considerable reduction in TRAs, raising a model wherein Aire functions predominantly as a direct transcriptional activator of TRA genes, and reduced TRAs is the cause of autoimmunity in Aire-KO mice (Anderson et al. 2002). This story seems perspicuous and reasonable, but some questions remain. Kuroda et al. reported that although mRNA levels of α-fodrin in mTECs were not reduced, autoantibodies against this molecule were produced in their Aire-deficient mouse model (Kuroda et al. 2005). Another example came from the Aire-KO mice on NOD (non-obese diabetic) background that developed severe autoimmune pancreatitis attacking acinar cells in parallel with a production of autoantibodies against pancreas-specific protein disulfide isomerase (PDIp), despite that the expression of PDIp was retained in Aire-deficient mTECs (Niki et al. 2006). Although further study is required, it is possible that Aire-dependent TRA reduction may not be the sole factor for the breakdown of self-tolerance in Aire-KO mice. In this regard, the role of Aire in the maturation program of mTECs has been proposed (Matsumoto 2011). Interestingly, each TRA protein is expressed only in a few mTECs, considered to be 1–3% of total mTECs with ordered stochasticity (Derbinski et al. 2008). The complete expression of all TRAs by the total mTEC population must be owing to the summation of mosaic expression of TRAs by individual mTECs (Kadouri et al. 2020).

10.8.3 The Molecular Function of Aire

Aire protein is localized in the nucleus as the shape of nuclear dots, resembling promyelocytic leukemia (PML) nuclear bodies, but they were revealed largely not to be colocalized (Akiyoshi et al. 2004). Considering its localization and structure, the Aire protein appears to be a putative transcriptional regulator, consisting of two plant homeodomain-type zinc-fingers (PHD-fingers), a DNA-binding domain (SAND), and four nuclear receptor binding LXXLL motifs (Kumar et al. 2001). These structural and functional domains are well conserved across phyla (Saltis et al. 2008). Many studies argue about the transcriptional role in Aire, but several unique features that differ from conventional transcription factors have been reported. Aire is apparently involved in the regulation of its target loci in collaboration with lots of partner proteins, forming a large multimolecular complex (Mathis and Benoist 2009). For example, CREB-binding protein (CBP) was the first identified Aire's partner (Pitkanen et al. 2000). It has also been reported that Aire recruits p-TEFb for transcriptional elongation of target genes (Oven et al. 2007), followed by a study arguing that bromodomain-containing protein, Brd4, bridges Aire and p-TEFb (Yoshida et al. 2015). Furthermore, a broad screen for Aire-targeted coimmunoprecipitation followed by high-throughput mass spectrometry newly identified putative Aire-interacting proteins involved in multiple biological pathways, including nuclear transport, chromatin structure, binding to the transcription machinery, and pre-mRNA processing (Abramson et al. 2010).

Aire's extraordinary broad transcriptional effect seems to be achieved by activating ectopic transcription, not through specific recognition of TRA gene promoters or enhancer motif. Instead, Aire appears to bind to the repressive chromatin mark H3K4me0 with its PHD1 finger domain (Koh et al. 2008; Org et al. 2008), and release RNA polymerase II paused just downstream of transcriptional start site (TSS) (Giraud et al. 2012). Moreover, recent bioinformatics revealed that Aire-containing complexes are predominantly located on mTEC super-enhancers, which are chromatin stretches enclosing TSS of Aire-dependent genes (Bansal et al. 2017).

10.8.4 mTEC Heterogeneity Defined by the Single-Cell Approach

As described above, TECs have been divided into cTECs (EpCAM^{+}Ly51^{+}UEA1^{-}) and mTECs (EpCAM^{+}Ly51^{-}UEA1^{+}), histologically and cytologically. Referring to their ontogeny, the evidence regarding the bipotent progenitor cells that give rise to both mTEC and cTEC lineages is emerging in the fetal and early neonatal thymus (Bleul et al. 2006; Rossi et al. 2006), characterized by cTEC-like molecular markers (Baik et al. 2013; Ohigashi et al. 2013). In contrast, it is still controversial about the existence and molecular characteristics of corresponding progenitors in the adult thymus (Ulyanchenko et al. 2016; Wong et al. 2014).

Depending on their molecular characteristics, mTECs were previously categorized as mTEC^{low} (Aire^{-}CD80^{low}MHC-II^{low}) and mTEC^{high} (Aire^{+}CD80^{high}MHC-II^{high}). "Central tolerance" is primarily achieved by the effective expression and presentation of TRAs from mTEC^{high} to the developing thymocytes. mTEC^{high} are differentiated from a part of mTEC^{low} and require RANK and CD40 signals for the development (Akiyama et al. 2008). In comparison with mTEC^{high}, mTEC^{low} fraction appeared to contain multiple subsets as studied in the past several years: (i) developing stage of mTEC lineage (recently categorized as "Ccl21^{+} mTEC") and (ii) terminally differentiated stage of mTECs, called "post-Aire mTEC" (Nishikawa et al. 2010) or “corneocyte-like mTEC” (Kadouri et al. 2020) (Table 10.1). Post-Aire mTECs lose their nuclei as they form Hassall’s corpuscles. Notably, Aire-deficient mice have reduced numbers of Krt10^{+} post-Aire mTECs and impaired formation of Hassall’s corpuscles in their thymi, which suggests that Aire may control the differentiation program of mTECs (Matsumoto 2011; Yano et al. 2008).

Furthermore, recent high-throughput scRNA-seq revealed that TECs, especially mTECs, consist of more heterogeneous groups than previously appreciated (Bornstein et al. 2018; Dhalla et al. 2020; Miller et al. 2018; Miragaia et al. 2018). Bornstein et al. categorized mTECs into four subsets as follows: (i) mTEC I (Ccl21^{+} mTEC), (ii) mTEC II (previous “mTEC$^{\text{high}}$”), (iii) mTEC III (previous “post-Aire mTEC” or “corneocyte-like mTEC”), and (iv) a newly identified mTEC IV (called “thymic tuft cells”). The existence of the thymic tuft cells, which are considered to establish an immune microenvironment in the thymus, was simultaneously reported by two groups (Bornstein et al. 2018; Miller et al. 2018). Thymic tuft cells are remarkably similar to peripheral tuft cells existing at mucosal barriers in that they express canonical taste transduction pathway molecules and IL-25, whereas the expression of MHC-II and CD74 is characteristic to thymic tuft cells (Miller et al. 2018). Moreover, Dhalla et al. identified a “proliferating mTEC” cluster that exhibited upregulation of *Mki67* with *Aire*, but its biology is still controversial (Ishikawa et al. 2021).

Recently, two groups have reported scRNA-seq studies focusing on human TECs (Bautista et al. 2021; Park et al. 2020). Importantly, human TECs have been revealed to contain similar subsets to mouse TECs (i.e., mTEC I-IV), and the expression of

Table 10.1 mTEC clusters identified by scRNA-seq

Cluster			Molecular marker	Functional role
mTEC I	mTEC$^{\text{low}}$	PDPN^{+} jTEC (CCL21^{+} mTEC)	Pdpn, Ccl21, Sox4, Ascl1, Itgb4, Itga6	mTEC precursor? Recruitment of CCR7^{+} thymocytes
mTEC II	Mature mTEC$^{\text{high}}$	AIRE^{+} mTEC$^{\text{high}}$	Aire, Fezf2, CD80, CD86, MHC-II$^{\text{high}}$	Tolerance induction
mTEC III	Post-Aire mTEC, Hassall’s corpuscle(mTEC$^{\text{low}}$)	KRT10^{+} corneocyte-like mTEC	Krt1, Krt10, Spink5, Ivl	Tolerance induction?
mTEC IV	Thymic tuft cell(mTEC$^{\text{low}}$)	DCLK1+ Tuft mTEC	Pou2f3, Dclk1, L1cam, Trpm5	Promotion of type 2 immunity
[a]Neuroendocrine cell			NEUROD1, BEX1, CHGA	Peptide provision to other APCs?
[a]Myoid cell			MYOD1, MYOG, DES	
[a]Myelin cell			SOX10, MPZ	Unknown

[a] Specific for human TEC

TRA genes and APECED relevant genes are enriched in the *AIRE*-expressing mTEC^{high} cluster (Bautista et al. 2021). Bautista et al. also reported the existence of immature TECs, which express canonical TEC genes but lacking characteristic genes of cTECs and mTECs, from both datasets. Moreover, some unique TEC subsets that are specific to humans were identified. Both groups reported the existence of (i) *MYOD1* and *MYOG* expressing myoid cells, and (ii) *NEUROD1*, *NEUROG*, and *CHGA* expressing neuroendocrine cells. Bautista et al. further identified (iii) *SOX10* and *MPZ* expressing myelin cells. Interestingly, expressions of myasthenia gravis relevant genes (i.e., *CHRNA1*, *TTN*, and *MUSK*) were predominantly found in the myoid, and neuroendocrine subsets (Bautista et al. 2021). It evokes the possibility that these unique AIRE^{-} populations also participate in the induction of immune tolerance, while these cells may not directly present antigens due to their low levels of MHC (HLA) expression. In summary, recent transcriptome analysis at single-cell resolution revealed that the thymus orchestrates the establishment of self-tolerance by the coordination of quite heterogenous TEC subsets, collaborating with unique transcriptional machineries.

10.9 Conclusion

In this chapter, we have described the recent advances in transcriptome analyses especially focusing on the bulk RNA-seq and scRNA-seq approaches that helped our understanding of the immune system more globally. In the last part of the chapter, we touched on how these techniques have now been bringing a new paradigm for self vs. non-self-discrimination in the thymus. The study on Aire deficiency, a monogenic autoimmune disease, has underscored the importance of the advent of new technologies to draw a whole picture of transcriptional control of the immune system. We are hoping that the complete picture of the transcripts of each immune cell type and the integration of this knowledge will pave the way to a comprehensive understanding of the immune system from a novel viewpoint.

References

Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4:e6098

Abramson J, Giraud M, Benoist C, Mathis D (2010) Aire's partners in the molecular control of immunological tolerance. Cell 140:123–135

Ackerman GA (1964) Histochemical differentiation during neutrophil development and maturation. Ann N Y Acad Sci 113:537–565

Ahonen P, Myllarniemi S, Sipila I, Perheentupa J (1990) Clinical variation of autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) in a series of 68 patients. N Engl J Med 322:1829–1836

Aizat WM, Ismail I, Noor NM (2018) Recent development in omics studies. Adv Exp Med Biol 1102:1–9

Akiyama T, Shimo Y, Yanai H, Qin J, Ohshima D, Maruyama Y, Asaumi Y, Kitazawa J, Takayanagi H, Penninger JM et al (2008) The tumor necrosis factor family receptors RANK and CD40 cooperatively establish the thymic medullary microenvironment and self-tolerance. Immunity 29:423–437

Akiyoshi H, Hatakeyama S, Pitkanen J, Mouri Y, Doucas V, Kudoh J, Tsurugaya K, Uchida D, Matsushima A, Oshikawa K et al (2004) Subcellular expression of autoimmune regulator (AIRE) is organized in a spatiotemporal manner. J Biol Chem 279:33984–33991

Amaya-Uribe L, Rojas M, Azizi G, Anaya J-M, Gershwin ME (2019) Primary immunodeficiency and autoimmunity: a comprehensive review. J Autoimmun 99:52–72

Amit I, Regev A, Hacohen N (2011) Strategies to discover regulatory circuits of the mammalian immune system. Nat Rev Immunol 11:873–880

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Nature Precedings

Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, Turley SJ, von Boehmer H, Bronson R, Dierich A, Benoist C, Mathis D (2002) Projection of an immunological self shadow within the thymus by the aire protein. Science 298:1395–1401

Baik S, Jenkinson EJ, Lane PJ, Anderson G, Jenkinson WE (2013) Generation of both cortical and Aire(+) medullary thymic epithelial compartments from CD205(+) progenitors. Eur J Immunol 43:589–594

Baird AE, Soper SA, Pullagurla SR, Adamski MG (2015) Recent and near-future advances in nucleic acid-based diagnosis of stroke. Expert Rev Mol Diagn 15:665–679

Bansal K, Yoshida H, Benoist C, Mathis D (2017) The transcriptional regulator Aire binds to and activates super-enhancers. Nat Immunol 18:263–273

Bautista JL, Cramer NT, Miller CN, Chavez J, Berrios DI, Byrnes LE, Germino J, Ntranos V, Sneddon JB, Burt TD et al (2021) Single-cell transcriptional profiling of human thymic stroma uncovers novel cellular heterogeneity in the thymic medulla. Nat Commun 12:1096

Bengtsson M, Ståhlberg A, Rorsman P, Kubista M (2005) Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Res 15:1388–1392

Besnard M, Padonou F, Provin N, Giraud M, Guillonneau C (2021) AIRE deficiency, from preclinical models to human APECED disease. Dis Model Mech 14:dmm046359

Bleul CC, Corbeaux T, Reuter A, Fisch P, Monting JS, Boehm T (2006) Formation of a functional thymus initiated by a postnatal epithelial progenitor cell. Nature 441:992–996

Bornstein C, Nevo S, Giladi A, Kadouri N, Pouzolles M, Gerbe F, David E, Machado A, Chuprin A, Toth B et al (2018) Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells. Nature 559:622–626

Brady G, Barbara M, Iscove NN (1990) Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. Methods Mol Cell Biol 2:17–25

Break TJ, Oikonomou V, Dutzan N, Desai JV, Swidergall M, Freiwald T, Chauss D, Harrison OJ, Alejo J, Williams DW et al (2021) Aberrant type 1 immunity drives susceptibility to mucosal fungal infections. Science 371:eaay5731

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11:94

Burton KJ, Pimentel G, Zangger N, Vionnet N, Drai J, McTernan PG, Pralong FP, Delorenzi M, Vergères G (2018) Modulation of the peripheral blood transcriptome by the ingestion of probiotic yoghurt and acidified milk in healthy, young men. PLoS One 13:e0192947

Chaplin DD (2010) Overview of the immune response. J Allergy Clin Immunol 125:S3–S23

Chattopadhyay PK, Gierahn TM, Roederer M, Love JC (2014) Single-cell technologies for monitoring immune systems. Nat Immunol 15:128–135

Chaussabel D (2015) Assessment of immune status using blood transcriptomics and potential implications for global health. Semin Immunol 27:58–66

Chou C, Li MO (2018) Tissue-resident lymphocytes across innate and adaptive lineages. Front Immunol 9:2104

Corkum CP, Ings DP, Burgess C, Karwowska S, Kroll W, Michalak TI (2015) Immune cell subsets and their gene expression profiles from human PBMC isolated by vacutainer Cell Preparation Tube (CPT™) and standard density gradient. BMC Immunol 16:48

Daniloski Z, Jordan TX, Wessels H-H, Hoagland DA, Kasela S, Legut M, Maniatis S, Mimitou EP, Lu L, Geller E et al (2021) Identification of required host factors for SARS-CoV-2 infection in human cells. Cell 184:92–105.e116

Davenport EE, Burnham KL, Radhakrishnan J, Humburg P, Hutton P, Mills TC, Rautanen A, Gordon AC, Garrard C, Hill AVS et al (2016) Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. Lancet Respir Med 4:259–271

Derbinski J, Pinto S, Rosch S, Hexel K, Kyewski B (2008) Promiscuous gene expression patterns in single medullary thymic epithelial cells argue for a stochastic mechanism. Proc Natl Acad Sci U S A 105:657–662

Dhalla F, Baran-Gale J, Maio S, Chappell L, Hollander GA, Ponting CP (2020) Biologically indeterminate yet ordered promiscuous gene expression in single medullary thymic epithelial cells. EMBO J 39:e101828

Dumeaux V, Olsen KS, Nuel G, Paulssen RH, Børresen-Dale A-L, Lund E (2010) Deciphering normal blood gene expression variation – the NOWAC postgenome study. PLoS Genet 6:e1000873

Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P (1992) Analysis of gene expression in single live neurons. Proc Natl Acad Sci U S A 89:3010–3014

Engel P, Boumsell L, Balderas R, Bensussan A, Gattei V, Horejsi V, Jin B-Q, Malavasi F, Mortari F, Schwartz-Albiez R et al (2015) CD nomenclature 2015: human leukocyte differentiation antigen workshops as a driving force in immunology. J Immunol (Baltimore, Md : 1950) 195:4555–4563

Engleman EG, Benike CJ, Grumet FC, Evans RL (1981) Activation of human T lymphocyte subsets: helper and suppressor/cytotoxic T cells recognize and respond to distinct histocompatibility antigens. J Immunol (Baltimore, Md : 1950) 127:2124–2129

Finnish-German APECED Consortium (1997) An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. Nat Genet 17:399–403

Fridman WH, Pagès F, Sautès-Fridman C, Galon J (2012) The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer 12:298–306

Fugmann SD, Lee AI, Shockett PE, Villey IJ, Schatz DG (2000) The RAG proteins and V(D)J recombination: complexes, ends, and transposition. Annu Rev Immunol 18:495–527

Giraud M, Yoshida H, Abramson J, Rahl PB, Young RA, Mathis D, Benoist C (2012) Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. Proc Natl Acad Sci U S A 109:535–540

Golumbeanu M, Cristinelli S, Rato S, Munoz M, Cavassini M, Beerenwinkel N, Ciuffi A (2018) Single-cell RNA-seq reveals transcriptional heterogeneity in latent and reactivated HIV-infected cells. Cell Rep 23:942–950

Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, Kang B, Liu Z, Jin L, Xing R et al (2018) Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med 24:978–985

Hajdu SI (2003) A note from history: the discovery of blood cells. Ann Clin Lab Sci 33:237–238

Hardy RR, Hayakawa K (2001) B cell development pathways. Annu Rev Immunol 19:595–621

Harrington LE, Hatton RD, Mangan PR, Turner H, Murphy TL, Murphy KM, Weaver CT (2005) Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. Nat Immunol 6:1123–1132

Hayakawa K, Hardy RR, Parks DR, Herzenberg LA (1983) The "Ly-1 B" cell subpopulation in normal immunodefective, and autoimmune mice. J Exp Med 157:202–218

Homuth G, Wahl S, Müller C, Schurmann C, Mäder U, Blankenberg S, Carstensen M, Dörr M, Endlich K, Englbrecht C et al (2015) Extensive alterations of the whole-blood transcriptome are associated with body mass index: results of an mRNA profiling study involving two large population-based cohorts. BMC Med Genet 8:65

Ishikawa T, Akiyama N, Akiyama T (2021) In pursuit of adult progenitors of thymic epithelial cells. Front Immunol 12:621824

Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, Linnarsson S (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res 21:1160–1167

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types, vol 343. Science (New York, N.Y.), pp 776–779

Jayasinghe SN (2020) Reimagining flow cytometric cell sorting. Adv Biosyst 4:e2000019

Jiang Y, Li Y, Zhu B (2015) T-cell exhaustion in the tumor microenvironment. Cell Death Dis 6:e1792

Kadouri N, Nevo S, Goldfarb Y, Abramson J (2020) Thymic epithelial cell heterogeneity: TEC by TEC. Nat Rev Immunol 20:239–253

Karamitros D, Stoilova B, Aboukhalil Z, Hamey F, Reinisch A, Samitsch M, Quek L, Otto G, Repapi E, Doondeea J et al (2018) Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. Nat Immunol 19:85–97

Karsten SL, Kudo LC, Bragin AJ (2011) Use of peripheral blood transcriptome biomarkers for epilepsy prediction. Neurosci Lett 497:213–217

Kaufmann SHE (2017) Emil von Behring: translational medicine at the dawn of immunology. Nat Rev Immunol 17:341–343

Kisand K, Boe Wolff AS, Podkrajsek KT, Tserel L, Link M, Kisand KV, Ersvaer E, Perheentupa J, Erichsen MM, Bratanic N et al (2010) Chronic mucocutaneous candidiasis in APECED or thymoma patients correlates with autoimmunity to Th17-associated cytokines. J Exp Med 207:299–308

Klein CA, Seidl S, Petat-Dutter K, Offner S, Geigl JB, Schmidt-Kittler O, Wendler N, Passlick B, Huber RM, Schlimok G et al (2002) Combined transcriptome and genome analysis of single micrometastatic cells. Nat Biotechnol 20:387–392

Klein L, Kyewski B, Allen PM, Hogquist KA (2014) Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nat Rev Immunol 14:377–391

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201

Koh AS, Kuo AJ, Park SY, Cheung P, Abramson J, Bua D, Carney D, Shoelson SE, Gozani O, Kingston RE et al (2008) Aire employs a histone-binding module to mediate immunological tolerance, linking chromatin regulation with organ-specific autoimmunity. Proc Natl Acad Sci U S A 105:15878–15883

Koumakis L (2020) Deep learning models in genomics; are we there yet? Comput Struct Biotechnol J 18:1466–1473

Kumar PG, Laloraya M, Wang CY, Ruan QG, Davoodi-Semiromi A, Kao KJ, She JX (2001) The autoimmune regulator (AIRE) is a DNA-binding protein. J Biol Chem 276:41357–41364

Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, Yamada RG, Ueda HR, Saitou M (2006) An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Res 34:e42

Kuroda N, Mitani T, Takeda N, Ishimaru N, Arakaki R, Hayashi Y, Bando Y, Izumi K, Takahashi T, Nomura T et al (2005) Development of autoimmunity against transcriptionally unrepressed target antigen in the thymus of Aire-deficient mice. J Immunol 174:1862–1870

Kyewski B, Klein L (2006) A central role for central tolerance. Annu Rev Immunol 24:571–606

Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, Keren-Shaul H, Mildner A, Winter D, Jung S et al (2014) Immunogenetics. Chromatin state dynamics during blood formation, vol 345. Science (New York, N.Y.), pp 943–949

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214

Maecker HT, McCoy JP, Nussenblatt R (2012) Standardizing immunophenotyping for the human immunology project. Nat Rev Immunol 12:191–200

Mahata B, Zhang X, Kolodziejczyk AA, Proserpio V, Haim-Vilmovsky L, Taylor AE, Hebenstreit D, Dingler FA, Moignard V, Göttgens B et al (2014) Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Rep 7:1130–1142

Mathis D, Benoist C (2009) Aire. Annu Rev Immunol 27:287–312

Matsumoto M (2011) Contrasting models for the roles of Aire in the differentiation program of epithelial cells in the thymic medulla. Eur J Immunol 41:12–17

Matsumoto M, Rodrigues PM, Sousa L, Tsuneyama K, Matsumoto M, Alves NL (2019) The ins and outs of thymic epithelial cell differentiation and function. In: Passos GA (ed) Thymus transcriptome and cell biology. Springer, pp 35–66

McCue ME, McCoy AM (2017) The scope of big data in one medicine: unprecedented opportunities and challenges. Front Vet Sci 4:194

Mello VDF, Kolehmanien M, Schwab U, Pulkkinen L, Uusitupa M (2012) Gene expression of peripheral blood mononuclear cells as a tool in dietary intervention studies: what do we know so far? Mol Nutr Food Res 56:1160–1172

Miao Y-L, Xiao Y-L, Du Y, Duan L-P (2013) Gene expression profiles in peripheral blood mononuclear cells of ulcerative colitis patients. World J Gastroenterol 19:3339–3346

Miller CN, Proekt I, von Moltke J, Wells KL, Rajpurkar AR, Wang H, Rattay K, Khan IS, Metzger TC, Pollack JL et al (2018) Thymic tuft cells promote an IL-4-enriched medulla and shape thymocyte development. Nature 559:627–631

Miragaia RJ, Zhang X, Gomes T, Svensson V, Ilicic T, Henriksson J, Kar G, Lonnberg T (2018) Single-cell RNA-sequencing resolves self-antigen expression during mTEC development. Sci Rep 8:685

Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, Burdin N, Visan L, Ceccarelli M, Poidinger M et al (2019) RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep 26:1627–1640.e1627

Mosmann TR, Cherwinski H, Bond MW, Giedlin MA, Coffman RL (1986) Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. J Immunol (Baltimore, Md : 1950) 136:2348–2357

Mostafavi S, Yoshida H, Moodley D, LeBoité H, Rothamel K, Raj T, Ye CJ, Chevrier N, Zhang S-Y, Feng T et al (2016) Parsing the interferon transcriptional network and its disease associations. Cell 164:564–578

Myhre AG, Halonen M, Eskelin P, Ekwall O, Hedstrand H, Rorsman F, Kampe O, Husebye ES (2001) Autoimmune polyendocrine syndrome type 1 (APS I) in Norway. Clin Endocrinol 54:211–217

Nagamine K, Peterson P, Scott HS, Kudoh J, Minoshima S, Heino M, Krohn KJ, Lalioti MD, Mullis PE, Antonarakis SE et al (1997) Positional cloning of the APECED gene. Nat Genet 17:393–398

Niki S, Oshikawa K, Mouri Y, Hirota F, Matsushima A, Yano M, Han H, Bando Y, Izumi K, Matsumoto M et al (2006) Alteration of intra-pancreatic target-organ specificity by abrogation of Aire in NOD mice. J Clin Invest 116:1292–1301

Nikolich-Zugich J, Slifka MK, Messaoudi I (2004) The many important facets of T-cell repertoire diversity. Nat Rev Immunol 4:123–132

Nishikawa Y, Hirota F, Yano M, Kitajima H, Miyazaki J, Kawamoto H, Mouri Y, Matsumoto M (2010) Biphasic Aire expression in early embryos and in medullary thymic epithelial cells before end-stage terminal differentiation. J Exp Med 207:963–971

Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, Habib N, Yosef N, Chang CY, Shay T et al (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144:296–309

Ohigashi I, Zuklys S, Sakata M, Mayer CE, Zhanybekova S, Murata S, Tanaka K, Hollander GA, Takahama Y (2013) Aire-expressing thymic medullary epithelial cells originate from beta5t-expressing progenitor cells. Proc Natl Acad Sci U S A 110:9885–9890

Org T, Chignola F, Hetenyi C, Gaetani M, Rebane A, Liiv I, Maran U, Mollica L, Bottomley MJ, Musco G, Peterson P (2008) The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression. EMBO Rep 9:370–376

Oven I, Brdickova N, Kohoutek J, Vaupotic T, Narat M, Peterlin BM (2007) AIRE recruits P-TEFb for transcriptional elongation of target genes in medullary thymic epithelial cells. Mol Cell Biol 27:8815–8823

Packer D (2021) The history of the antibody as a tool. Acta Histochem 123:151710

Park H, Li Z, Yang XO, Chang SH, Nurieva R, Wang Y-H, Wang Y, Hood L, Zhu Z, Tian Q, Dong C (2005) A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. Nat Immunol 6:1133–1141

Park JE, Botting RA, Dominguez Conde C, Popescu DM, Lavaert M, Kunz DJ, Goh I, Stephenson E, Ragazzini R, Tuck E et al (2020) A cell atlas of human thymic development defines T cell repertoire formation. Science 367:eaay3224

Pauken KE, Sammons MA, Odorizzi PM, Manne S, Godec J, Khan O, Drake AM, Chen Z, Sen DR, Kurachi M et al (2016) Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. Science (New York, NY) 354:1160–1165

Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, Reinmaa E, Sutphin GL, Zhernakova A, Schramm K et al (2015) The transcriptional landscape of age in human peripheral blood. Nat Commun 6:8570

Peterson P, Peltonen L (2005) Autoimmune polyendocrinopathy syndrome type 1 (APS1) and AIRE gene: new views on molecular basis of autoimmunity. J Autoimmun 25(Suppl):49–55

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods 10:1096–1098

Pitkanen J, Doucas V, Sternsdorf T, Nakajima T, Aratani S, Jensen K, Will H, Vahamurto P, Ollila J, Vihinen M et al (2000) The autoimmune regulator protein has transcriptional transactivating properties and interacts with the common coactivator CREB-binding protein. J Biol Chem 275:16802–16809

Proserpio V, Mahata B (2016) Single-cell technologies to study the immune system. Immunology 147:133–140

Puel A, Doffinger R, Natividad A, Chrabieh M, Barcenas-Morales G, Picard C, Cobat A, Ouachee-Chardin M, Toulon A, Bustamante J et al (2010) Autoantibodies against IL-17A, IL-17F, and IL-22 in patients with chronic mucocutaneous candidiasis and autoimmune polyendocrine syndrome type I. J Exp Med 207:291–297

Qin Y, Yalamanchili HK, Qin J, Yan B, Wang J (2015) The current status and challenges in computational analysis of genomic big data. Big Data Res 2:12–18

Rankin LC, Artis D (2018) Beyond host defense: emerging functions of the immune system in regulating complex tissue physiology. Cell 173:554–567

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25

Rossi SW, Jenkinson WE, Anderson G, Jenkinson EJ (2006) Clonal analysis reveals a common progenitor for thymic cortical and medullary epithelium. Nature 441:988–991

Rothenberg EV (2014) Transcriptional control of early T and B cell developmental choices. Annu Rev Immunol 32:283–321

Rouse BT, Sehrawat S (2010) Immunity and immunopathology to viruses: what decides the outcome? Nat Rev Immunol 10:514–526

Roy AL (2019) Transcriptional regulation in the immune system: one cell at a time. Front Immunol 10:1355

Saelens W, Cannoodt R, Todorov H, Saeys Y (2019) A comparison of single-cell trajectory inference methods. Nat Biotechnol 37:547–554

Sakaguchi S, Sakaguchi N, Asano M, Itoh M, Toda M (1995) Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. J Immunol (Baltimore, Md : 1950) 155:1151–1164

Sakaguchi S, Mikami N, Wing JB, Tanaka A, Ichiyama K, Ohkura N (2020) Regulatory T cells and human disease. Annu Rev Immunol 38:541–566

Saltis M, Criscitiello MF, Ohta Y, Keefe M, Trede NS, Goitsuka R, Flajnik MF (2008) Evolutionarily conserved and divergent regions of the autoimmune regulator (Aire) gene: a comparative analysis. Immunogenetics 60:105–114

Sánchez-Ramón S, Bermúdez A, González-Granado LI, Rodríguez-Gallego C, Sastre A, Soler-Palacín P (2019) Primary and secondary immunodeficiency diseases in oncohaematology: warning signs, diagnosis, and management. Front Immunol 10:586

Schmidt M, Hopp L, Arakelyan A, Kirsten H, Engel C, Wirkner K, Krohn K, Burkhardt R, Thiery J, Loeffler M et al (2020) The human blood transcriptome in a large population cohort and its relation to aging and health. Front Big Data 3:548873

Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B, Altay G, Greenbaum JA, McVicker G et al (2018) Impact of genetic polymorphisms on human immune cell gene expression. Cell 175:1701–1715.e1716

Seumois G, Vijayanand P (2019) Single-cell analysis to understand the diversity of immune cell types that drive disease pathogenesis. J Allergy Clin Immunol 144:1150–1153

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D et al (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498:236–240

Shen-Orr SS, Gaujoux R (2013) Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol 25:571–578

Smale ST, Fisher AG (2002) Chromatin structure and gene regulation in the immune system. Annu Rev Immunol 20:427–462

Sohn E (2017) Diagnosis: frontiers in blood testing. Nature 549:S16–S18

Stein M, Keshav S, Harris N, Gordon S (1992) Interleukin 4 potently enhances murine macrophage mannose receptor activity: a marker of alternative immunologic macrophage activation. J Exp Med 176:287–292

Stewart CA, Gay CM, Xi Y, Sivajothi S, Sivakamasundari V, Fujimoto J, Bolisetty M, Hartsfield PM, Balasubramaniyan V, Chalishazar MD et al (2020) Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. Nat Can 1:423–436

Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA (2017) Single-cell transcriptomics to explore the immune system in health and disease. Science (New York, NY) 358:58–63

Talal N (1973) Lymphocyte heterogeneity and function. Arthritis Rheum 16:422–425

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A et al (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382

Uhlen M, Karlsson MJ, Zhong W, Tebani A, Pou C, Mikes J, Lakshmikanth T, Forsström B, Edfors F, Odeberg J et al (2019) A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science (New York, NY) 366:eaax9198

Ulyanchenko S, O'Neill KE, Medley T, Farley AM, Vaidya HJ, Cook AM, Blair NF, Blackburn CC (2016) Identification of a bipotent epithelial progenitor population in the adult thymus. Cell Rep 14:2819–2832

Veenstra TD (2021) Omics in systems biology: current progress and future outlook. Proteomics 21:e2000235

Wong K, Lister NL, Barsanti M, Lim JM, Hammett MV, Khong DM, Siatskas C, Gray DH, Boyd RL, Chidgey AP (2014) Multilineage potential and self-renewal define an epithelial progenitor cell population in the adult thymus. Cell Rep 8:1198–1209

Xie X, Liu M, Zhang Y, Wang B, Zhu C, Wang C, Li Q, Huo Y, Guo J, Xu C et al (2021) Single-cell transcriptomic landscape of human blood cells. Natl Sci Rev 8:nwaa180

Yano M, Kuroda N, Han H, Meguro-Horike M, Nishikawa Y, Kiyonari H, Maemura K, Yanagawa Y, Obata K, Takahashi S et al (2008) Aire controls the differentiation program of thymic epithelial cells in the medulla for the establishment of self-tolerance. J Exp Med 205:2827–2838

Yoshida H, Bansal K, Schaefer U, Chapman T, Rioja I, Proekt I, Anderson MS, Prinjha RK, Tarakhovsky A, Benoist C, Mathis D (2015) Brd4 bridges the transcriptional regulators, Aire and P-TEFb, to promote elongation of peripheral-tissue antigen transcripts in thymic stromal cells. Proc Natl Acad Sci U S A 112:E4448–E4457

Yoshida H, Lareau CA, Ramirez RN, Rose SA, Maier B, Wroblewska A, Desland F, Chudnovskiy A, Mortha A, Dominguez C et al (2019) The cis-regulatory atlas of the mouse immune system. Cell 176(897-912):e820

Part II
Transcriptome in Disease

Chapter 11
Transcriptome Profiling in Autoimmune Diseases

Cristhianna V. A. Collares and Eduardo A. Donadi

The principal function of the immune system is related to the protection of the organism against invasion of pathogens and restore tissue integrity. However, the immune system may fail to perform its major task in immunodeficiency states, usually associated with the inability of its major components to respond/fight pathogens, and in autoimmune diseases, associated with the failure to distinguish between the self and non-self. The violation of the tolerance to self-antigens is the basis of autoimmune diseases, which will be focused on this chapter.

Autoimmune diseases are a group of different inflammatory disorders, characterized by systemic or localized inflammation, usually leading to cell or tissue destruction. Systemic autoimmune disorders encompass diseases caused by a fault in the immune system, primarily associated with the failure to recognize self-antigens by tolerance loss and cross-reactivity between pathogen and self-antigens. In these processes, the chronic overreactivity of B and T cells induces unsafe signals to cells or tissues, impairing their function. The uncontrolled cell proliferation may also promote chronic inflammation (Kamradt and Mitchison 2001; Matzinger 2002).

The knowledge on the pathogenesis of autoimmune diseases has not yet been fully elucidated. Studies suggest that some specific gene variants or shared gene variants are observed autoimmune diseases and, similarly, pathogenic mechanisms may be shared among these disorders (Zhernakova et al. 2009; Cho and Gregersen 2011). Although many genetic *loci* have been described for autoimmune diseases, additional elements have been identified and associated with pathogenesis of these disorders.

Approximately 0.1–1% of general population develops autoimmune diseases during life. In a recent study, researchers have reported that almost 20 million

C. V. A. Collares · E. A. Donadi (✉)
Department of Medicine, Division of Clinical Immunology, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil
e-mail: eadonadi@fmrp.usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_11

Americans are affected by autoimmune diseases (Rose 2016). Considering first-degree relatives, this incidence increases five times, and in monozygotic twins, this rate increases more than five times. Thus, the risk is increased with increasing genetic similarity to an affected individual. However, the highest autoimmune disease concordance rate (higher than 20–30%) is observed among monozygotic twins, showing that additional genetic and non-genetic factors may play a role (Mackay 2009). Besides genetics, some autoimmune diseases are influenced by hormonal factors and affect more women than men, i.e., systemic lupus erythematosus, that affects 80–90% more women than men, and the peak of its incidence occurs during childbearing ages (Straub 2007). Indeed, the role of hormonal factors corroborates an evolutionary point of view, in which the female gender of most vertebrates has an incubation time to passively offer immunity to the offspring (immunological priming) and during pregnancy and breastfeeding (Lin et al. 2016; Moulton 2018). If evolution has shaped immunological priming, an undesirable effect of a stronger immune response in the trend for the development of autoimmune diseases. Several lines of evidence indicate the participation of sex hormones on the regulation of both innate and adaptive (humoral and cell-mediated) immune responses. Noteworthy, sexual hormones may act in the development, homeostasis, and gene expression, in addition to controlling the signaling processes of T and B lymphocytes, influencing gene expression profiles (Klein and Flanagan 2016; Ortona et al. 2016; Trombetta et al. 2017; Edwards et al. 2018).

In addition to genes and hormones, recent studies regarding the influence of infectious agents (particularly Arbovirus and Coronavirus) on the modification of the immune response have shed light into the understanding of the pathogenesis of autoimmune disorders. The outbreak of Zika, Chikungunya, and Dengue viruses, which have reached Brazil and other countries in the 2015–2016 epidemics, has produced neurological and rheumatological complications that are remarkably like multiple sclerosis, neuromyelitis optica, Guillain-Barré syndrome, and others. Besides cross-reaction between host and virus antigens, the viruses themselves have been detected in the cerebrospinal fluid (Almeida et al. 2021), putatively influencing gene expression profiles. The recent pandemic caused by the infection by SARS-CoV2 has also produced several alterations on the leukocyte function and of the inflammatory cytokine profile (Lei et al. 2020), producing autoimmune manifestation exacerbations (Peeters et al. 2020).

Among the several autoimmune disorders, in this revision we will devote special attention to three conditions, an endocrinologic (type 1 diabetes mellitus), a rheumatologic (systemic lupus erythematosus), and a neurologic (multiple sclerosis) autoimmune disorders. These three diseases have been prominently investigated in the medical-scientific scenario, particularly emphasizing their transcriptional profiles, and this knowledge may contribute for a better understanding of disease pathogenesis, disease morbidity, development of new drugs and diagnostic tools, and the unraveling of biomarkers for early diagnosis and treatment.

11.1 Type 1 Diabetes Mellitus

Diabetes affects more than 300 million of adults in worldwide population, and this rate will be increase to 440 million by 2030 (Shaw et al. 2010). Type 1 diabetes (T1D) is usually associated with an autoimmune pathogenesis and accounts for approximately 10% of all cases of diabetes, affecting individuals under the age of 30, but can also be manifested later (Geenen et al. 2010). The incidence and prevalence of T1D have grown worldwide, particularly in developed countries, increasing 2–3% per year (Maahs et al. 2010; Mayer-Davis et al. 2017). In the United States, the increased incidence of T1D is observed among young people under 15 years of age, and especially in children under 5 years old (Chobot et al. 2017). Although susceptibility to T1D has been primarily attributed to histocompatibility genes (HLA), most of the patients have no familiar history of the disease and no HLA susceptibility genes, indicating a combinatorial influence of environmental and behavioral factors (DiMeglio et al. 2018).

T1D is caused by cellular and humoral autoimmune responses specifically against pancreatic beta cells, producing a substantial decrease of insulin production (Battaglia 2014). It is estimated that 80–95% of pancreatic beta cells are destroyed when T1D is diagnosed, impairing the diagnosis at the early asymptomatic stages of the autoimmune attack. Since the direct access to pancreas is difficult, disease diagnosis is primarily clinical, and the search for islet autoantibodies may have important function as serological marker for the disease (Achenbach et al. 2004). Autoantibodies against islet cells antigens (ICAs), glutamic acid decarboxylase (GAD), islet antigens (IAA), and protein tyrosine phosphatase-like protein IA-2 (IA-2A) can be used for the prediction and progression of T1D (Achenbach et al. 2004). In recent years, other biomarkers have been used like autoantibodies against the zinc efflux transporter ZnT8 (Herold et al. 2009; Wenzlau et al. 2007). Most newly diagnosed T1D patients have measurable antibodies against insulin, GAD, IA-2, ZnT8, and tetraspanin-7 (McLaughlin et al. 2016). It has also been described that seroconversion of two or more autoantibodies is necessary for the development of T1D in almost 85% of cases before the age of 20 years (Ziegler et al. 2013). Despite the presence of these serological biomarkers, the discovery of novel biomarkers to diagnose T1D before the complete islet cell destruction is a major goal. The early prediction is still missing due to insufficient predictive power of the individual risk factors (Purohit and She 2008).

Several genes and gene regions distributed throughout the genome have been associated with T1D in population, family, and linkage studies. T1D has been shown to be slightly more common in men than in women (Diaz-Valencia et al. 2015). The strongest genetic susceptibility contribution comes from the human leucocyte antigen complex (HLA) region at chromosome 6p21 (IDDM1), responsible for up to 40–50% of T1DM susceptibility, and from the insulin gene (*INS*) region (IDDM2) (Pugliese and Miceli 2002). Although it is still unknown how some HLA haplotypes can interfere with the susceptibility or resistance or development of T1D, some studies have shown that in 50% of cases, susceptibility has been associated with the

*HLA-DRB1**03:01-*DQA1**05:01-*DQB1**02:01 (DR3/DQ2) and *HLA-DRB1**04:01-*DQA1**03:01-*DQB1**03:01 (DR4-DQ8) haplotypes, and these haplotypes have been more prevalent in White individuals. In contrast, protection against the development of T1D has been associated with the *HLA-DRB1**15:01-*DQA1**01:02-*DQB1**06:02 (DR15-DQ6) haplotype (Noble 2015). In addition, more than 60 non-HLA *loci* have been identified as being relevant for the development of T1D in genome-wide association studies, including immune system variants, such as the expression of the insulin thymus gene, regulation of T cell activation, and viral responses (Noble 2015).

The loss of tolerance to self-antigens in T1D can be summarized combining the genetic features and the immunoregulatory imbalance observed in T1D patients, encompassing: (i) decreased expression of insulin in thymus, impairing lymphocyte education during the negative selection in the medullary region; (ii) the molecules encoded by the genes *HLA-DQA1**05:01, *HLA-DQB1**03:02, and *DQB1**02:01 and *HLA-DRB1**03/*04 mediate the presentation of autoantigens and this feature is associated with the development of anti-GAD, insulin (IAA), islet antigen 2 (IA2A), and ISS autoantibodies; (ii) deficient immunoregulation, mediated by specific surface and intracellular molecules, including IL-2, IL-2RA, IL-2RB, CTLA-4, PTPN-2, and PTPN-22; (iii) decreased number of T regulatory cells; (iv) decreased function of the molecules involved on innate immune response; (v) decreased cell regulation by invariant natural killer-like T (iNKT) cells (Cipolletta et al. 2005; Sia 2006; Li et al. 2007; Chentoufi et al. 2008; Knip and Siljander 2008; McDevitt and Unanue 2008; Tisch and Wang 2008; Karumuthil-Melethil et al. 2008; Todd 2010; Pociot et al. 2010; Buschard 2011; Novak and Lehuen 2011).

Diabetes mellitus is one of the most studied diseases in terms of genetic association and of the transcriptome profile of disease subtypes. Pathogenic features observed in T1D patients have been used to characterize the gene expression profile and to identify novel biomarkers, for instance: (i) among the more than 40 *loci* associated with probable contributors to T1D (Barrett et al. 2009; Nerup et al. 2009; Plagnol et al. 2011), most of them modulate the immune system (Concannon et al. 2009); (ii) chromatin remodeling may simultaneously downregulate several inflammatory genes and upregulate many genes responsible for a set of cellular functions, including glucose homeostasis, and other signaling and metabolic pathways (Jayaraman et al. 2013); (iii) proinflammatory cytokines or double-stranded sRNA (by-product resulting from viral infection) can contribute for T1D pathogenesis, modifying the expression of the differentially expressed genes observed in human pancreatic islets (Eizirik et al. 2009; Moore et al. 2009; Colli et al. 2010; Eizirik et al. 2012); (iv) cytokines may regulate more than 3000 genes associated with inflammation, innate immune response and apoptosis, and cytokine and chemokine genes (CCL2, CCL5, CCL3, CXCL9, CXCL10, CXCL11, IL-6, and IL-8), and these genes are induced in human islet cells. Among them, the CCL2 and CXCL10 molecules attract macrophages and may be involved in the recruitment of immune cells at the beginning of insulitis (Eizirik et al. 2012); and (v) the highly expressed CXCL10 chemokine is regulated by 16-5p miRNA, and the binding of this miRNA to the *CXCL10* transcript can act on T1D cell proliferation and apoptosis in T1D patients (Gao and Zhao 2020).

Differential gene expression profiles have also been reported for humans and animal models (Grinberg-Bleyer et al. 2010; Fornari et al. 2011), using peripheral blood mononuclear cells (Rassi et al. 2008; Han et al. 2011; Collares et al. 2013b), pancreatic beta cells (Planas et al. 2010), and in whole peripheral blood cells (Reynier et al. 2010) obtained from T1D patients.

To unveil shared and specific differences among autoimmune and non-autoimmune diabetes, several studies have been conducted using peripheral blood lymphocytes, primarily applying the microarray technology. In this context, Collares et al. (2013b) evaluated the transcriptome of diabetic patients, including T1D, type 2 diabetes (T2D), and gestational diabetes (GDM) patients. The results revealed that the overall gene expression profile is characteristic for each group of diabetic patients and that gene expression profile of GDM was closer to T1D than to T2D. An in silico analysis showed that the similarities observed in the transcriptional profile of GDM and T1D were due to the role of genes associated with inflammation (Evangelista et al. 2014). The higher expression of these genes in some T1D and GDM patients seems to influence the global gene expression pattern of diabetic patients. Indeed, several important molecular mechanisms identified in this cluster account for an intricate array of the inflammation pathways. Using the DAVID database (http://david.abcc.ncifcrf.gov/) it is possible to obtain functionality of these genes and the involvement of modulated genes in different biological functions. In a meta-analysis evaluating the transcription profiles of T1D, T2D, and GDM, it was observed that induced genes were grouped into five major groups of biological function: (i) development of multicellular organism (20.3%), (ii) signal transduction (17.9%), (iii) stress response (12.2%), (iv) cell differentiation (10.7%), and (v) processes the immune system (6.8%). The repressed genes were clustered into three main biological processes: (i) regulation of metabolic processes (30%), (ii) biosynthetic processes (26.9%), and (iii) transcriptional processes (22%) (Fig. 11.1). Performing a more restrictive analysis, considering only modulated genes that showed fold change ≥ 2 for all comparisons of the three groups of diabetes, we observed 10 most significant genes, seven of them were induced in GDM and T1D and repressed in T2D and three genes were repressed in GDM and T1D and induced in T2D. The chromosomal location of these genes is shown in Fig. 11.2.

Pancreatic beta cell gene features have been studied by real-time PCR, microarrays (Kutlu et al. 2009; Dorrell et al. 2011), and, more recently, using next-generation sequencing (NGS) of RNA (Mutz et al. 2013). Among the studies involving NGS, it is important to highlight that the genome-wide association study (GWAS) from human diabetic populations (pancreatic islets) confirmed more than 60% of previously identified genes in T1D (Eizirik et al. 2012). The major contribution of GWAS to T1D was reviewed in detail, expanding the six susceptibility genomic regions for T1D to almost 60 regions, emphasizing the importance to identify and characterize intronic variants and intergenic regions. Additionally, a major advance in this area was the large-scale sequencing, mainly studies involving scRNA-seq, which allowed understanding the role of coding and non-coding RNAs and their influences on T1D

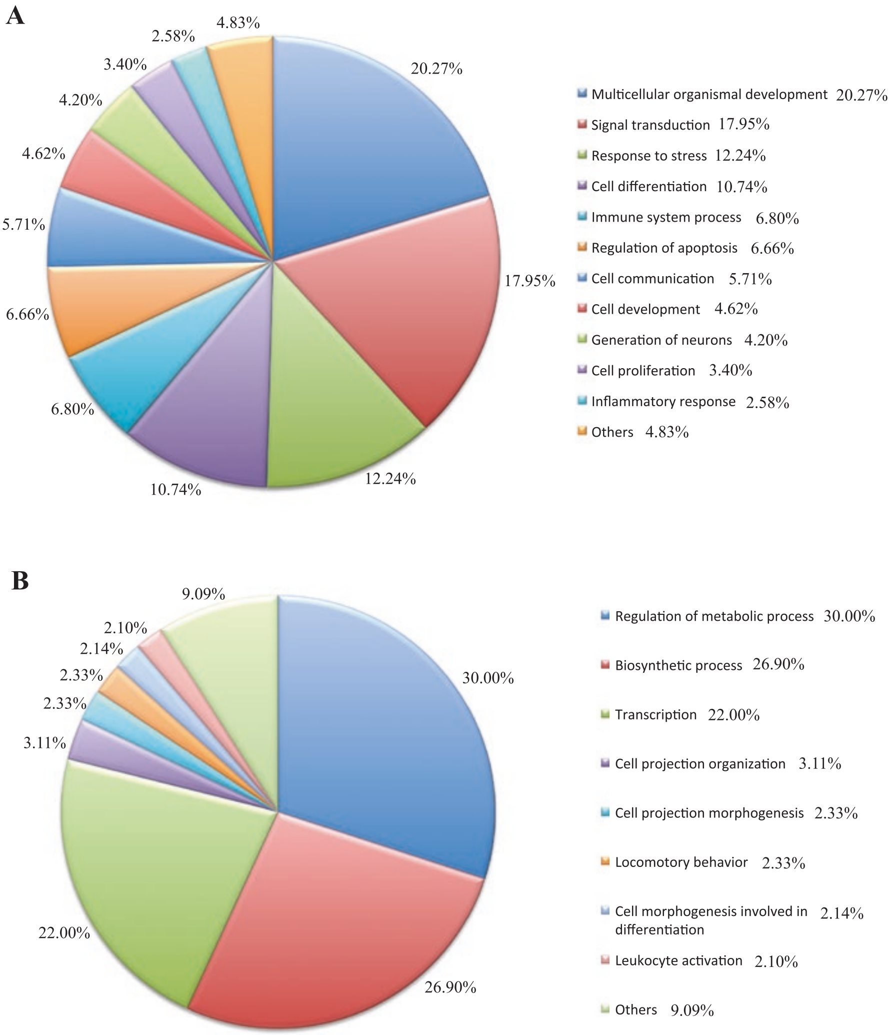


Fig. 11.1 Biological function of the significant and differentially expressed genes (3747 transcripts) modulated after comparing GDM, T1DM, and T2DM. Panel A illustrates the upregulated transcripts (from GDM to T1D to T2D), which were clustered into five groups according to their biological functions: (i) development of multicellular organism (20.3%), (ii) signal transduction (17.9%), (iii) stress response (12.2%), (iv) cell differentiation (10.7%), and (v) immune system processes (6,8%). Panel B shows the downregulated transcripts clustered into three main groups: (i) regulation of metabolic processes (30%), (ii) biosynthetic processes (26.9%), and (iii) transcriptional processes (22%)

development (Bakay et al. 2019). The non-coding RNAs have been identified as important or even major regulators of gene expression and include small microRNA (miRNA) and long noncoding RNA (lncRNA) (Mattick and Makunin 2006; Ponting et al. 2009). Approximately 200 dysregulated miRNAs in several tissues, cells, and

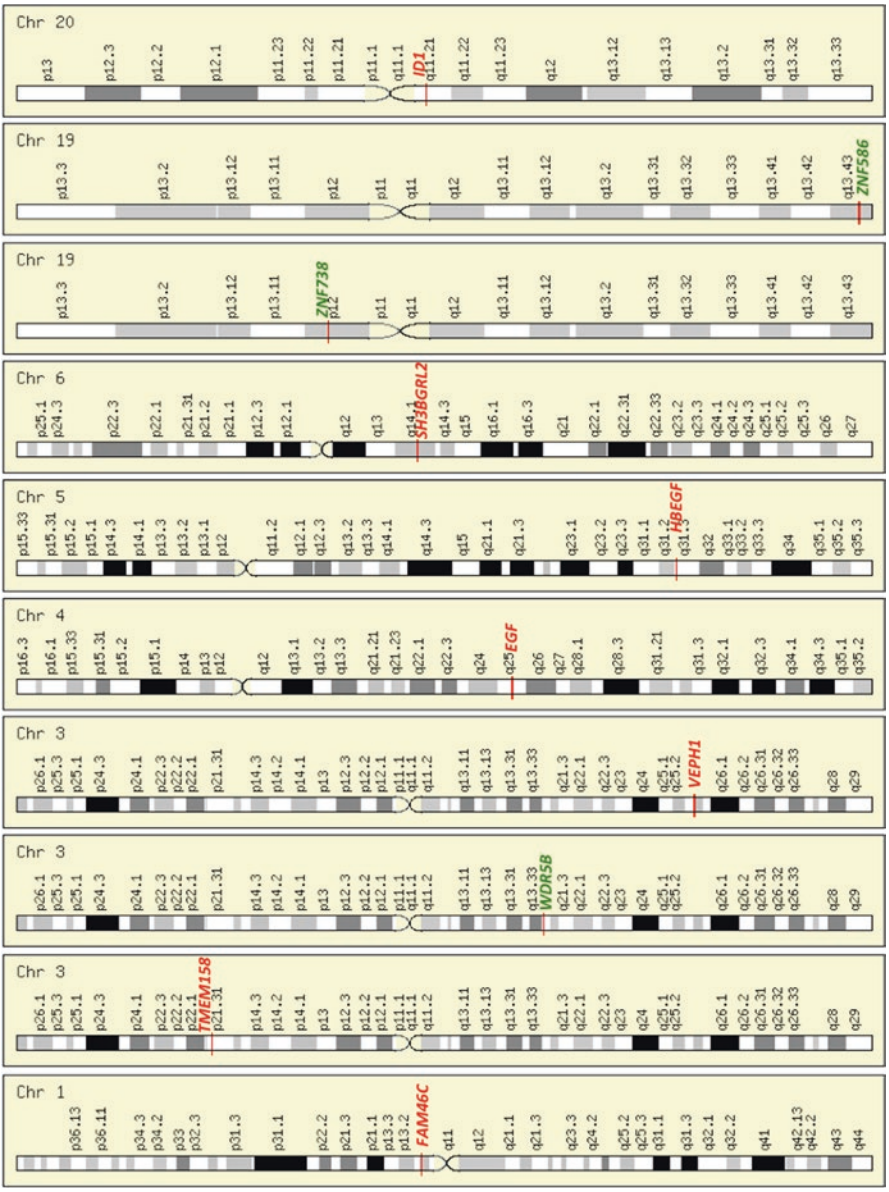

Chr 20
p13 p12.3 p12.2 p12.1 p11.23 p11.22 p11.21 p11.1 q11.1 q11.21 q11.22 q11.23 q12 q13.12 q13.13 q13.2 q13.31 q13.32 q13.33
ID1
Chr 19
p13.3 p13.2 p13.12 p13.11 p12 p11 q11 q12 q13.11 q13.12 q13.2 q13.31 q13.32 q13.33 q13.41 q13.42 q13.43
ZNF586
Chr 19
p13.3 p13.2 p13.12 p13.11 p12 p11 q11 q12 q13.11 q13.12 q13.2 q13.31 q13.32 q13.33 q13.41 q13.42 q13.43
ZNF738
Chr 6
p25.1 p24.3 p22.3 p22.1 p21.31 p21.2 p21.1 p12.3 p12.1 q12 q13 q14.1 q14.3 q15 q16.1 q16.3 q21 q22.1 q22.31 q22.33 q23.2 q23.3 q24.1 q24.2 q24.3 q25.1 q25.2 q25.3 q26 q27
SH3BGRL2
Chr 5
p15.33 p15.31 p15.2 p15.1 p14.3 p14.1 p13.3 p13.2 p13.1 p12 q11.2 q12.1 q12.3 q13.2 q13.3 q14.1 q14.3 q15 q21.1 q21.3 q23.1 q23.2 q23.3 q31.1 q31.2 q31.3 q32 q33.1 q33.2 q33.3 q34 q35.1 q35.2 q35.3
HBEGF
Chr 4
p16.3 p16.1 p15.33 p15.31 p15.2 p15.1 p14 p13 p12 q12 q13.1 q13.2 q13.3 q21.21 q21.23 q22.1 q22.3 q24 q25 q26 q27 q28.1 q28.3 q31.21 q31.3 q32.1 q32.3 q34.1 q34.3 q35.1 q35.2
EGF
Chr 3
p26.1 p25.3 p25.1 p24.3 p24.1 p22.3 p22.2 p22.1 p21.31 p14.3 p14.2 p14.1 p13 p12.3 p12.2 p12.1 p11.1 q11.1 q11.2 q13.11 q13.13 q13.31 q13.33 q21.3 q22.1 q22.3 q23 q24 q25.1 q25.2 q26.1 q26.2 q26.31 q26.32 q26.33 q28 q29
VEPH1
Chr 3
p26.1 p25.3 p25.1 p24.3 p24.1 p22.3 p22.2 p22.1 p21.31 p14.3 p14.2 p14.1 p13 p12.3 p12.2 p12.1 p11.1 q11.1 q11.2 q13.11 q13.13 q13.31 q13.33 q21.3 q22.1 q22.3 q23 q24 q25.1 q25.2 q26.1 q26.2 q26.31 q26.32 q26.33 q28 q29
WDR5B
Chr 3
p26.1 p25.3 p25.1 p24.3 p24.1 p22.3 p22.2 p22.1 p21.31 p14.3 p14.2 p14.1 p13 p12.3 p12.2 p12.1 p11.1 q11.1 q11.2 q13.11 q13.13 q13.31 q13.33 q21.3 q22.1 q22.3 q23 q24 q25.1 q25.2 q26.1 q26.2 q26.31 q26.32 q26.33 q28 q29
TMEM158
Chr 1
p36.13 p36.11 p34.3 p34.2 p33 p32.3 p31.3 p31.1 p22.2 p21.3 p21.1 p13.3 p13.2 q11 q12 q21.1 q21.3 q23.3 q24.2 q25.2 q25.3 q31.1 q31.3 q32.1 q32.2 q41 q42.13 q42.2 q43 q44
FAM46C

Fig. 11.2 Chromosomal location of the ten differentially expressed genes selected from the meta-analysis among GDM, T1D, and T2D and exhibiting fold change ≥2

blood (serum and plasma) have been identified in both human and murine T1D sample Assmann et al. 2017). Some miRNAs are key players in pancreatic development and homeostasis. Among them, miR-375 has been considered a key miRNA that regulates B-cell insulin secretion and hence overall glucose homeostasis

(Dumortier and Van Obberghen 2012). Moreover, miR-375 was reported as putative circulating biomarker for beta-cell injury (Erener et al. 2013), since it has been responsible for insulin gene expression and secretion (El Ouaamari et al. 2008; Poy et al. 2004). Overall, miRNAs play regulatory roles in many biological processes associated with diabetes, including adipocyte differentiation, metabolic integration, insulin resistance, and appetite regulation (Krützfeldt and Stoffel 2006). The role of miRNAs in diabetes has been associated with several pathogenic features. For example, miR-410, miR-200a, and miR-130a regulate secretion of insulin in response to stimulatory levels of glucose, and overexpression of miR-410 enhances the levels of glucose-stimulated insulin secretion (Hennessy et al. 2010). The miR-200 and the miR-7 families are constitutively repressed in beta-cells (Latreille et al. 2014; Belgardt et al. 2015). MiR-30d is upregulated in pancreatic beta-cells and collaborates for the increased insulin gene expression (Tang et al. 2009a) and miR-9 acts in the fine-tuning of glucose metabolism (Plaisance et al. 2006).

In experimental diabetes, the miR-142-3p, miR-142-5p, and miR-155 present in NOD (non-obese diabetic) mice (observed also in human T-cell exosomes) favored apoptosis of beta cells, and when inactivated, protected mice from the development of T1D. The islets of the protected mice have a higher level of insulin, a lower rate of insulitis, and inflammation. This is because the exosomes of T lymphocytes trigger apoptosis and the expression of genes involved in chemokine signaling, such as CCL2, CCL7, and CXCL10, in beta cells. The induction of these genes can promote the recruitment of immune cells and exacerbate the death of beta cells during the autoimmune attack (Guay et al. 2019).

The comparisons between the transcript profile of the immunologic and non-immunologic types of diabetes revealed the mRNA/miRNA signatures of T1D, T2D, and GDM patients, pinpointing some miRNAs shared among the three types of diabetes, miRNAs specific for each type of diabetes, and identified non-described miRNAs associated with each type of diabetes (Collares et al. 2013a). Nine miRNAs were shared among the three types of diabetes, including hsa-miR-126, hsa-miR-144, hsa-miR-27a, hsa-miR-29b, hsa-miR-1307, hsa-miR-142-3p, hsa-miR-142-5p, hsa-miR-199a-5p, and hsa-miR-342-3p, and suggested that these miRNAs are associated with diabetes *per se*. For T1D, some miRNAs were primarily observed in T1D and the only miRNA that linked to glucose metabolism was let-7f, which was previously suggested as a potential therapy for T2D (Frost and Olson 2011). Another biomarker that has already been suggested for early detection of beta-cell death and predicting T1D development was the methylation patterns of circulating DNA (Lehmann-Werman et al. 2016). A NGS study evaluating T1D and T2D patients showed specific miRNAs involved in diabets complication: diabetic nephropathy (miR-9-5p, miR1249-3p, miR-409-5p, miR12271-5p, miR-501-3p, miR-193a-5p, and miR-148a-5p); diabetic neuropathy (miR-873-5p, miR-125a-5p, miR145-3p, and miR-99b-5p); diabetic retinopathy (miR-409-5p, miR-1271-5p, miR143-3p, and miR-199a-5p). In addiction, this study showed miRNAs involved in more than one diabetc complications: miR-193b-3p, miR-101-5p, miR-486-5p, miR-382-5p, miR-144-5p, and miR-145-3p. Most of the miRNAs differentially expressed in T1D and T2D patients targeted genes (*GSK3B, FZD4, BCL2, PRKAA1,* and *NFAT5*) and gene families (*SMAD, AKT, CBL, MAPK, TGF-beta,* and *FOXO*) associated with diabetic complications (Massaro et al. 2019).

Concluding, several genes and transcripts have already been described in autoimmune diabetes; many of than were associated with increased chance to develop T1D with increased morbidity, complications, and mortality. Although linkage studies have been associated with transcriptome and microRNAs and, most recently proteomes, many issues are still studied to unveil the intricate mechanisms associated with the development of diabetes. The understanding of the association between susceptibility genes and their differential transcript profiles may contribute to the development of new drugs, novel strategies for diabetes care, including early diagnosis (before there is destruction of most of the pancreatic beta cells), avoiding autoimmune destruction of pancreatic beta cells, or even preventing the development of T1D.

11.2 Systemic Lupus Erythematous

Systemic lupus erythematosus (SLE) is a systemic autoimmune disorder characterized by the presence of high amounts of circulating immune complexes, leading to tissue/organ damage because of persistent tissue inflammation (Shaikh et al. 2017; Kaul et al. 2016). Among SLE patients, women are nine times more affected than men, suggesting that gender-related factors are crucial for disease development (Schwartzman-Morris and Putterman 2012; Weckerle and Niewold 2011). The highest SLE prevalence (241/100.00) and incidence (23.2/100,000 people/year) rates are observed in North America, and the lowest in Africa and Ukraine (0.3/100,000 people/year) (Rees et al. 2017). The etiology of SLE remains uncertain and disease pathogenesis has been associated with the interaction of genetic, epigenetic, and environmental factors (Shaikh et al. 2017; Kaul et al. 2016; Liu and Davidson 2012). Exposure to viruses and bacterial infections, and ultra-violet radiation are known to trigger SLE (Doria et al. 2008). SLE involves the activation of the innate and adaptive immune response and is usually considered a severe and potentially life-threatening disease, which may represent a therapeutic challenge because of its heterogeneous organ manifestations. The central pathogenic features of SLE encompass T- and B-cells abnormalities that lead to autoantibody production. Innate immune cells produce type 1 interferon (IFN) that has a central role in systemic autoimmunity and in the activation of B and T cells. Autoantibodies produced by B-cells stimulate dendritic cell IFN production, combining the role of innate and adaptive immune system responses (Kiefer et al. 2012).

In terms of the genetic risk for developing SLE, only 10-20% of the cases can be explained by heritability, and genetic variability of individuals has a smaller contribution (Moser et al. 2009). The great challenge of the studies on lupus is the identification of gene variants that are, in 90% of the cases, in non-coding, intronic or intergenic regions (Moser et al. 2009; Deng and Tsao 2010; Guerra et al. 2012; Costa et al. 2013; Kilpinen and Dermitzakis 2012). Although: (i) the concordance rate of SLE in monozygotic twins is 24–57% and in dizygotic twins or full siblings is 2–5% (Jarvinen et al. 1992; Deapen et al. 1992; Ghodke-Puranik and Niewold 2015), (ii) no genomic differences between monozygotic twins are observed, evidencing the role of epigenomic and gene expression variations (Furukawa et al. 2013), (iii) environmental factors mediate epigenetic effects, i.e., DNA/RNA methylation and histone modification, these findings corroborate the complex bases of SLE heritability, emphasizing the key role of environmental factors in disease development (Alarcón-Segovia et al. 2005; Kilpinen and Dermitzakis 2012; Costa et al. 2013).

Genome-wide association and linkage studies have revealed more than 90 *loci* associated with SLE susceptibility (Kaul et al. 2016; Graham et al. 2009; Niewold 2015; Langefeld et al. 2017) and over 40 genes have been associated with pathways associated with immune system regulation, tissue response to injury, endothelial function, and others (Moser et al. 2009; Deng and Tsao 2010; Guerra et al. 2012). Linkage and genome-wide association studies have indicated genetic risk factors associated with the HLA region (Gough and Simmonds 2007; Morris et al. 2014; Armstrong et al. 2014), and with other genes including *IRF5, STAT4, BLK, TNFAIP3, TNIP1, FCGR2B,* and *TNFSF13* (Koga et al. 2011; Morris et al. 2016; Alarcón-Riquelme et al. 2016). Noteworthy, genes associated with SLE susceptibility genes are also involved in other autoimmune diseases, emphasizing the genetic polymorphism of cytokine genes that may differentially control lymphocyte activity. Several cytokines (IL-1, IL-6, IL-10, TNF-alpha, among others) play an important role in SLE disease activity (Asanuma et al. 2006; McCarthy et al. 2014; Cigni et al. 2014). According to this idea, cytokine signatures can be defined depending on the disease activity: (i) a "susceptibility signature" is observed in patients in clinical remission, (ii) an "activity signature" is related to genes associated with immune cell metabolism and protein synthesis and proliferation, and (iii) a "severe signature" is related to active nephritis (Panousis et al. 2019). In the context of nephritis, a recent study showed evidence of genetic risk shared between SLE and lupus nephritis, especially in patients younger than 18 years, pointing that SLE susceptibility *loci* are related to the development of proliferative lupus nephritis (Webber et al. 2020).

The group of the interferon (IFN) cytokines deserves special attention in SLE. Since peripheral blood mononuclear cells (PBMC) are reporters of the ongoing tissue/cell/organ damage occurring elsewhere, these cells have been used to evaluate the transcript profiles in autoimmune disorders. Using PBMC, the IFN signature is very characteristic and more prominent in patients with more active and severe form of SLE. Induced genes have been associated with type I interferon signaling (Crow and Kirou 2004; Ghodke-Puranik and Niewold 2013; Crow 2014), with emphasis on interferon alpha (IFN-alfa) (Blanco et al. 2001; Niewold 2011; Weckerle et al. 2011). Microarray analysis showed overexpression of IFN-alfa-regulated genes in SLE patients, including the type 1 IFN signature (Crow and Wohlgemuth 2003; Niewold et al. 2007; Panousis et al. 2019). Moreover, it has shown that viral infection treated with IFN-alfa may contribute to *de novo* SLE development that disappears when treatment is discontinued (Niewold and Swedler 2005; Ronnblom et al. 1990).

PBMC transcript profiles may also differentiate clinical manifestations of SLE, differentiating neuropsychiatric from non-neuropsychiatric patients (Sandrin-Garcia et al. 2012). Besides PBMC, subpopulations of purified cells may exhibit a unique pattern of gene expression. A specific signature has been identified in SLE T CD4+ cells, involving IFN transcripts, and most differentially expressed genes in these cells had promoter sequences presenting targets for the interferon regulatory factor (IRF) -3 and -7 (Lyons et al. 2010; Li et al. 2010). The involvement of IFN and IFN-induced genes also appear when evaluating target organ or tissue. In the transcriptome of the synovial membrane, the comparisons between SLE patients with rheumatoid arthritis or osteoarthritis show upregulation of IFN induced genes and repression of genes involved in extracellular matrix homeostasis (Toukap et al. 2007).

The bone marrow is a central lymphoid organ with hematopoietic and immunoregulatory function and exhibits a variety of histopathological abnormalities in SLE, and the evaluation of bone marrow may be more informative than PBMC (Voulgarelis et al. 2006). The differential gene expression of the bone marrow, using the microarray analysis, shows a clear differentiation between active from inactive SLE (Nakou et al. 2008), revealing pathways related to cellular growth, cell survival, and immune reactions, as important factors associated with SLE pathogenesis (Nakou et al. 2010). Additionally, the transcriptome analysis of SLE platelets reveals increased expression of genes encoding cytokines, chemokines, and proteins involved in apoptosis, and overexpression of type I IFN-regulated genes in comparison to controls (Lood et al. 2010). These studies confirm the type I INF-related genes expression profile in platelets from SLE patients, as well as its related proteins. Regardless of the biological material studied, several lines of evidence show that IFN is extremely important for SLE, since its expression is induced in SLE tissues and cells. Approximately 60% of patients exhibit increased expression of genes induced by type 1 IFN, which is directly associated with the disease activity (Baechler et al. 2003; Bennett et al. 2003; Han et al. 2003; Kirou et al. 2004; Kirou et al. 2005; Feng et al. 2006), and with signaling pathways induced by type 1 IFN (Yao et al. 2009). Other interferon-induced genes are also modulated in SLE patients, including OAS2 (2′5′-oligoadenylate synthetase 2) (Grammatikos et al. 2014), and interferon regulatory factor 5 (*IRF5*), for which specific polymorphisms may confer susceptibility to SLE (Stone et al. 2013). The gene expression profile studies of granulocytes show differential expression of genes involved in cell apoptosis and motility, and also show repression of genes related to DNA repair, differential expression of genes involved in cell apoptosis and motility (Baechler et al. 2003; Han et al. 2003; Rus et al. 2004; Maas et al. 2005; Lee et al. 2011).

Many transcription factors were shown to be crucial for immune system and their differences in the expression and activity may imply in discovering novel biomarkers in some diseases, including SLE. Comparing levels of transcription factors in PBMC of SLE patients, Sui et al. (2012) found 92 differentially expressed transcription factors and indicated activator protein-1 (AP-1), Pbx1, and myocyte enhancer factor-2 (MEF-2) as candidates involved in pathogenesis of SLE and new diagnosis biomarker for this disease. The transcription factor FOXO1 was also related to SLE, which was downregulated in PBMCs from SLE and rheumatoid arthritis patients (Kuo and Lin 2007). The transcript family FOXO involves transcription factors that play an important role in controlling lymphocyte activation and proliferation. A member of nuclear factor (NF)-kB/Rel family of transcription factors, c-Rel, was found in higher levels in PBMCs from SLE patients (Burgos et al. 2000). Since cytokines are produced by T-help cell 1 (Th1) and 2 (Th2), probably transcription factors related to T-help cells must have an important role in SLE (Foster and Kelley 1999). The principal transcription factors for differentiation of Th1 and Th2 are T-bet and GATA-3, which were found to be up or downregulated in SLE patients, respectively (Chan et al. 2006; Lit et al. 2007). Other transcription factors, including AP-1, NF-kB, and IRF5, increase STAT-4 expression (Remoli et al. 2007) and are important for type 1 IFN receptor signaling. Moreover, the IRF5 is mediator of Toll-like receptor-triggered expression of proinflammatory cytokines such as type 1 IFN and TNF-alfa (Kawai and Akira 2006).

Regarding the control of gene expression within the context of SLE, the levels of RNA may be controlled by epigenetic mechanism including microRNAs, which usually acts by degradation of target mRNA or inhibiting its translation. Many studies have reported miRNAs deregulation in SLE and more than 42 differentially expressed microRNAs were detected in PBMCs from SLE patients, and some of them were pinpointed as biomarker candidates. It has been demonstrated that miRNA deregulation is implicated in different systemic autoimmune diseases (Stagakis et al. 2011). MiR-21 acts partly through inhibition of *PDCD4* (selective protein translation inhibitor of genes involved in immune responses) and it was found upregulated in T- and B-cells (Stagakis et al. 2011) and in CD4 T cells (Pan et al. 2010) of SLE patients comparing to control group, suggesting it as possible biomarker for SLE.

An increased expression of miR-224 (Lu et al. 2013), miR-148a (Pan et al. 2010), miR-15 (Yuan et al. 2012), miR-142-3p and miR-181a (Carlsen et al. 2013), miR-189, miR-61, miR-78, miR-21, miR-142-3p, miR-342, miR-299-3p, miR-198, and miR-298 (Dai et al. 2007) has been described in patients and animal models. However, many studies have shown downregulation of different microRNAs in SLE patients, including: miR-146a (Tang et al. 2009b), miR-145 in T cells (Lu et al. 2013), miR-155 in serum and urine from SLE patients (Wang et al. 2010), miR-181a in pediatric patients (Lashine et al. 2011), miR-19b and miR-20a in monocytes (Teruel et al. 2011), miR-125a (Zhao et al. 2010), miR-17, miR-20a, miR-106a, miR-92a, and miR-203 in the circulation (Carlsen et al. 2013), and miR-196a, miR-17-5p, miR-409-3p, miR-141, miR-383, miR-112, and miR-184 in PBMCs (Dai et al. 2007). In addition, a circulating miRNA (miR-125a 3p) was related to disease activity (Wang et al. 2011, 2012).

In the context of SLE, some specific knowledge about some miRNAs and their target genes may help to develop new drugs and to propose novel diagnostic tools. For example, the decreased expression of miR-145 and induction of its target protein activator of transcription-1 (STAT-1) seem to be associated with lupus nephritis, and may contribute to the immunopathogenesis of SLE (Lu et al. 2013). MiR-146a, which targets STAT1 and IRF5 in innate immune cells and is negative regulator of type 1 IFN and TLR7 signaling pathways, was described as repressed compared to controls (Tang et al. 2009b). Some reports show miRNAs, such as miR-148a, miR-126, miR-21, and miR-29b, are downregulated, targeting DNA methyltransferase-1 expression and, thus, contributing to global hypomethylation observed in SLE (Deng et al. 2001; Pan et al. 2010; Zhao et al. 2011; Layer et al. 2003; Qin et al. 2013). MiR-125a is involved in inflammatory chemokine pathway and contributes to higher expression of RANTES, an inflammatory chemokine, indicating that this miR can be used as a novel target for SLE treatment (Zhao et al. 2010). MiRNA-142 and miR-31 have already been reported in T cells of SLE patients, acting on cytokine expression (Ding et al. 2012; Fan et al. 2012). MiR-146a and miR-241-3p/5p have been described as altered by mycophenolic acid, followed by reduction of autoreactive lupus T cells, a finding that suggests that these miRNAs may serve as prediction biomarkers to drugs (Tang et al. 2015).

11.3 Multiple Sclerosis

Multiple sclerosis (MS) is a common, severe, chronic inflammatory autoimmune, and demyelinating disease of the central nervous system (CNS), associated with an immune reaction against myelin proteins. The disease primarily affects the white matter, in which autoreactive T cells attack the myelin-oligodendrocyte complex (Noseworthy et al. 2000). Generally, it begins at the third and fourth decades of life, affects more women than men (3:1), and is more common in developed countries of the Northern hemispheres (Weinshenker 1994; Orton et al. 2006). Approximately 80% of MS patients have relapsing-remitting MS forms (Lublin and Reingold 1996).

The etiology of MS is still unknown; however, evidence indicates a multifactorial and complex nature, where genetic and environmental factors may influence their onset (Noseworthy et al. 2000). Evidence pinpoints for polygenic susceptibility and for multiple environmental triggering factors (Poo 2001), including Epstein-Barr virus (EBV) infection, smoking, obesity, and vitamin D deficiency (Ramagopalan et al. 2010; Mokry et al. 2016; Sintzel et al. 2018).

The role of autoimmunity becomes clear by the presence of autoreactive T cells for myelin components of CNS and peripheral blood of MS patients. It is believed that T lymphocytes are activated at lymph nodes in the periphery and bind to receptors on endothelial cell, continuing to cross the blood–brain barrier into the interstitial matrix (Karpuj et al. 1997). Activation of T cells induces the release of cytokines, propitiating sensitized lymphocytes to have access to the CNS through the blood–brain barrier, stimulating chemotaxis. The recruitment of inflammatory cells and leakage of plasma proteins into the CNS trigger a series of mechanisms responsible for myelin damage. The main pathological feature in MS is the plaque, a well-demarcated white matter injury, histologically characterized by inflammation, T cells and macrophage infiltration, demyelination and gliosis, and axonal loss (Lucchinetti et al. 1998).

Current knowledge of MS allows the formation of the concept of circulating T-cell receptor-selected T cells in MS and that CD8+ T cells may be essential in the pathophysiology of the disease. The study of abnormalities of blood T cells in MS may contribute to better understanding of the disease and the discovery of new drugs against MS (Laplaud et al. 2004). The study of monozygotic twins and the observation of the differential MS manifestations have suggested the influence of environmental factors (Mumford et al. 1994; Sadovnick et al. 1996; Willer et al. 2003; Nielsen et al. 2005; Islam et al. 2006; Chitnis 2007; Oksenberg et al. 2008; Harirchian et al. 2018). In genetically associated cases, the genetic component is valued at higher relative risk of siblings of affected individual presenting the same disease, and there is also a higher concordance rate in monozygotic than in dizygotic twins (Sadovnick and Ebers 1993; Willer et al. 2003). Considering MS, the most striking susceptibility genes are encoded at the major histocompatibility complex MHC, especially class II alleles (Dyment et al. 2004). The HLA-DR2 phenotype (*DRB1*15:01-DQB1*06:02*) has been described in different populations (Sadovnick and Ebers 1993; Epplen et al. 1997; Barcellos et al. 2003; Hollenbach and Oksenberg 2015). It is believed that there are specific standards of ethnic, environmental, or both association patterns, wherein *HLA-DRB1* alleles may have different behavior in different environmental contexts (Brum et al. 2007). Furthermore, the *DRB1*15* allele group was suggested as significant factor MS susceptibility and development (Kaimen-Maciel et al. 2009), as previously demonstrated in the Brazilian Caucasian population (Brum et al. 2007).

Many research studies of gene expression on MS have been performed using brain tissue from patients, and gene profile may be altered in acute, chronic, or silent lesions, or even normal tissue. The most important discoveries were related to a set of 16 genes related to autoimmunity, with seven of them associated with SLE and two associated with T1D (Tajouri et al. 2007).

GWAS studies revealed more than 150 single nucleotide polymorphisms in MS patients, many of them located in regulatory regions of genes related to the immune response (Dobson and Giovannoni 2019). Approximately 300 differentially expressed genes were detected in a study done in PBMCs of MS patients (Bomprezzi et al. 2003). Among them, overexpression of (i) platelet activating factor acetyl hydrolase (*PAFAH1B1*), a gene associated with brain development and chemoattraction during inflammation and allergy; (ii) tumor necrosis factor receptor (*TNFR* or *CD27)*, which is a co-stimulator for T cell activation and fundamental for immune response development; (iii) T cell receptor (*TCR*), crucial for T cell mediated immune response and it was associated with MS susceptibility (Beall et al. 1993); (iv) zeta chain associated protein kinase (*ZAP70*), gene responsible for TCR induced T cell activation (Chan et al. 1992); (v) interleukin 7 receptor (*IL7R*), involved in B and T cells activation. In the same study, several genes were repressed, such as: tissue inhibitor of metalloproteinase 1 (*TIMP1*), plasminogen activator inhibitor 1 (*SERPINE 1*), histone coding genes, and heat shock protein 70 (*HSP70)*. Additionally, the evaluation of T cells from MS patients stressed the importance of transcriptional regulation of NF-kB, which is responsible for regulating gene expression during MS relapse; deregulation of NF-kB on T cell transcriptome may be used as a molecular biomarker for clinical disease activity (Satoh et al. 2008).

An alternative study of the transcription profile of MS patients is the use of cerebrospinal fluid. Brynedal et al. (2010) investigated gene expression profile in leukocytes of CSF from MS patients and found *AIF1, MGC29506, POU2AF1, PLAUR,* and *TNFRSF17* as differentially expressed. A comparative study between MS patients at relapse and healthy controls showed the overexpression of genes involved in T and NK cell process, genes belonging to pathways involved in T-cell co-stimulation, activated T-cell proliferation, regulation of cell surface receptors, and NK-cell activation (Jernas et al. 2013). The authors also showed a decreased expression of genes associated with innate immunity, B-cell activation and immunoglobulin secretion, and T helper 2 responses in leukocytes of CSF, highlighting the *HMOX1* gene. The deletion of this gene was associated with enhanced demyelination (Chora et al. 2007). The induced genes were: (i) *EDN1*, associated with integrity of blood–brain barrier; (ii) *CXCL11* that is important for recruitment of T-cells to the CNS when disease activity is higher; and (iii) *CXCL13*, which may be important for the T-helper cell recruitment during relapses. Furthermore, certain *CXCL13* polymorphic sites associated with high levels of the chemokine are more frequent in patients with MS (Lindén et al. 2013). Cerebrospinal fluid was also used for the study of the hypothalamus–pituitary–adrenal (HPA) axis activity in MS, because of its association with disease progression and comorbid mood disorders. The activity of axis was determined by measuring cortisol in cerebrospinal fluid and the results revealed, in MS patients, low HPA axis activity and associated it with increased disease severity (Melief et al. 2013).

Additionally, in a comparative study between three neurodegenerative disorders (associated neurocognitive disorders, Alzheimer's disease, and multiple sclerosis) was observed the common overexpression of *BACE2*, gene previously associated with Alzheimer's disease (Holler et al. 2012) that codes for an amyloid-beta peptide (Borjabad and Volsky 2012). In the same study, among the repressed genes were the *GABRG2* (GABA receptor 2), impairing GABAergic neuron signal transmission and memory (Melzer et al. 2012). Observing T cell genes in whole blood of MS patients, Gandhi et al. (2010) showed overexpressed genes in MS patients comparing to control subjects, and most of them was expressed on cells from antigen presenting cell, suggesting that excessive T cell activity as a hallmark of disease.

Brain derived neurotrophic factor (BDNF) was suggested as neuroprotective factor for MS (Frota et al. 2009) and the overexpression of anti-inflammatory pathway, BNDF related neuroprotection, showed by overexpression of BDNF, BDNF upstream activator-TNK, and BDNF receptor NTRK3, was demonstrated during acute relapse (Gurevich and Achiron 2012). In addition, some transcript factors were described influencing MS disease, in special the YY1, which is related to processes that affect myelin protein generation (Berndt et al. 2001), immune response process (Guo et al. 2001; Guo et al. 2008), and viral replication (Oh and Broyles 2005) and are involved in differential gene expression in MS patients (Riveros et al. 2010).

Epigenetic mechanisms may alter gene expression and modulate the response to environmental factors, affecting MS morbidity. Three principal epigenetic mechanisms include: DNA methylation, histone modifications, and micro-RNA-mediated genetic silencing. Among them, miRNAs have been extensively evaluated for their influence on the manifestation of various autoimmune diseases, including MS. Several microRNAs were induced in different MS studies, some of them were just induced when compared to controls and others were associated with disease activity/severity, including: (i) miR-17-5p that acts on lipid kinases and regulates the development of lymphocyte (Lindberg et al. 2010); (ii) miR-326 (associated with Th17 cell profile) (Chen et al. 2018) was associated with disease severity (Du et al. 2009); (iii) miR-214 and miR-23a were present in active and inactive MS lesions, and in oligodendrocyte differentiation, suggesting their involvement in remyelination (Junker et al. 2009); (iv) miR-23a was observed in PBMCs from patients exhibiting the remissive/remittent disease subset (Ridolfi et al. 2013); (v) miR-338, miR-491, and miR-155 (also referred as miR-155-5p) were associated with more advanced stages of MS (Noorbakhsh et al. 2011); (vi) miR-155 was observed in peripheral blood monocytes and in myeloid cells from MS brain lesions (Moore et al. 2013); (vii) miR-145, miR-660, miR-939, and miR-223 were observed in PBMCs (Sondergaard et al. 2013; Ridolfi et al. 2013), blood (Keller et al. 2009; Cox et al. 2010), and T regulatory cells (De Santis et al. 2010); (viii) miR-34a, miR-142–3p, and miR-326 were detected in demyelinating plaques, suggesting that these miRNAs may contribute to disease pathogenesis (Junker et al. 2009; Mandolesi et al. 2017; Honardoost et al. 2014).

Considering the repressed miRNA in MS patients: (i) miR-219 and miR-338-5p were observed in inactive lesions, targeting genes responsible for the integrity of myelin (Junker et al. 2009); (ii) serum miR-15b and miR-223 that target genes implicated MS pathogenesis (Fenoglio et al. 2013); (iii) members of the mir-29 family observed in PBMCs from relapsing-remitting patients were associated with apoptotic processes and IFN feedback loops (Hecker et al. 2013); (iv) miR-20a-5p in whole blood of patients (Keller et al. 2014) that targets the *CDKN1A* gene, which collaborates in T cell activation and has been associated with systemic autoimmunity (Santiago-Raber et al. 2001); (v) miR-17 and miR-20a, which are related to control of immune function, are involved in T cell activation and are implicated in MS pathogenesis (Cox et al. 2010).

In conclusion, in this revision we highlighted three representative autoimmune diseases: T1D, SLE, and MS. In all of them, the non-genetic factors may have important role in the development of the disorders. The discovery of gene transcripts and miRNAs that are involved in the development of each one of these disorders is a major challenge to increase the understanding of their role on disease pathogenesis, which may be useful to develop new drugs and novel diagnostic/prognostic tools. Prevention may also be feasible when biomarkers (susceptibility genes and differentially expressed transcripts) are available, permitting the early detection of autoimmune disorders, ameliorating patient care.

References

Achenbach P, Warncke K, Reiter J et al (2004) Stratification of type 1 diabetes risk on the basis of islet autoantibody characteristics. Diabetes 53:384–392

Alarcón-Riquelme ME, Ziegler JT, Molineros J et al (2016) Genome-wide association study in an Amerindian ancestry population reveals novel systemic lupus erythematosus risk loci and the role of European admixture. Arthritis Rheum 68:932–943

Alarcón-Segovia D, Alarcón-Riquelme ME, Cardiel MH et al (2005) Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1,177 lupus patients from the GLADEL cohort. Arthritis Rheum 52:1138–1147

Almeida RS, Ferreira MLB, Sonon P et al (2021) Cytokines and soluble HLA-G levels in the acute and recovery phases of arbovirus-infected Brazilian patients exhibiting neurological complications. Front Immunol 12:582935

Armstrong D, Zidovetzki R, Alarcón-Riquelme M et al (2014) GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. Genes Immun 15:347–354

Asanuma Y, Chung CP, Oeser A et al (2006) Increased concentration of proatherogenic inflammatory cytokines in systemic lupus erythematosus: relationship to cardiovascular risk factors. J Rheumatol 33:539–545

Assmann TS, Recamonde-Mendoza M, De Souza BM et al (2017) MicroRNA expression profiles and type 1 diabetes mellitus: systematic review and bioinformatic analysis. Endocr Connect 6:773–790

Baechler EC, Batliwalla FM, Karypis G et al (2003) Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. PNAS 100:2610–2615

Bakay M, Pandey R, Grant SFA et al (2019) The genetic contribution to type 1 diabetes. Curr Diab Rep 19:116

Barcellos LF, Oksenberg JR, Begovich AB et al (2003) HLA-DR2 dose effect on susceptibility to multiple sclerosis and influence on disease course. Am J Hum Genet 72:710–716

Barrett JC, Clayton DG, Concannon P et al (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703–707

Battaglia M (2014) Neutrophils and type 1 autoimmune diabetes. Curr Opin Hematol 21:8–15

Beall SS, Biddison WE, McFarlin DE et al (1993) Susceptibility for multiple sclerosis is determined, in part, by inheritance of a 175-kb region of the TcR V beta chain locus and HLA class II genes. J Neuroimmunol 45:53–60

Belgardt BF, Ahmed K, Spranger M et al (2015) The microRNA-200 family regulates pancreatic beta cell survival in type 2 diabetes. Nat Med 21:619–627

Bennett L, Palucka AK, Arce E et al (2003) Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med 197:711–723

Berndt JA, Kim JG, Tosic M et al (2001) The transcriptional regulator Yin Yang 1 activates the myelin PLP gene. J Neurochem 77:935–942

Blanco P, Palucka AK, Gill M et al (2001) Induction of dendritic cell differentiation by IFN-alpha in systemic lupus erythematosus. Science 294:1540–1543

Bomprezzi R, Ringner M, Kim S et al (2003) Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. Hum Mol Genet 12:2191–2199

Borjabad A, Volsky DJ (2012) Common transcriptional signatures in brain tissue from patients with HIV-associated neurocognitive disorders, Alzheimer's disease, and multiple sclerosis. J NeuroImmune Pharmacol 7:914–926

Brum DG, Barreira AA, Louzada-Junior P et al (2007) Association of the HLA-DRB1*15 allele group and the DRB1*1501 and DRB1*1503 alleles with multiple sclerosis in White and Mulatto samples from Brazil. J Neuroimmunol 189:118–124

Brynedal B, Khademi M, Wallstrom E et al (2010) Gene expression profiling in multiple sclerosis: a disease of the central nervous system, but with relapses triggered in the periphery? Neurobiol Dis 37:613–621

Burgos P, Metz C, Bull P et al (2000) Increased expression of c-rel, from the NF-KB/Rel family, in T cells from patients with systemic lupus erythematosus. J Rheumatol 27:116–127

Buschard K (2011) What causes type 1 diabetes? Lessons from animal models. APMIS Suppl 119(132):1–19

Carlsen AL, Schetter AJ, Nielsen CT et al (2013) Circulating microRNA expression profiles associated with systemic lupus erythematosus. Arthritis Rheum 65:1324–1334

Chan AC, Iwashima M, Turck CW et al (1992) ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR zeta chain. Cell 71:649–662

Chan RW, Lai FM, Li EK et al (2006) Imbalance of Th1/Th2 transcription factors in patients with lupus nephritis. Rheumatology 45:951–957

Chen C, Zhou Y, Wang J et al (2018) Dysregulated microRNA involvement in multiple sclerosis by induction of T helper 17 cell differentiation. Front Immunol 9:1256

Chentoufi AA, Binder NR, Berka N et al (2008) Advances in type I diabetes associated tolerance mechanisms. Scand J Immunol 68:1–11

Chitnis T (2007) The role of CD4 T cells in the pathogenesis of multiple sclerosis. Int Rev Neurobiol 79:43–72

Cho JH, Gregersen PK (2011) Genomics and the multifactorial nature of human autoimmune disease. N Engl J Med 365:1612–1623

Chobot A, Polanska J, Brandt A et al (2017) Updated 24-year trend of type 1 diabetes incidence in children in Poland reveals a sinusoidal pattern and sustained increase. Diabet Med 34:1252–1258

Chora AA, Fontoura P, Cunha A et al (2007) Heme oxygenase-1 and carbon monoxide suppress autoimmune neuroinflammation. J Clin Invest 117:438–447

Cigni A, Pileri PV, Faedda R et al (2014) Interleukin 1, interleukin 6, interleukin 10, and tumor necrosis factor α in active and quiescent systemic lupus erythematosus. J Investig Med 62:825–829

Cipolletta C, Ryan KE, Hanna EV et al (2005) Activation of peripheral blood CD14+ monocytes occurs in diabetes. Diabetes 54:2779–2786

Collares CV, Evangelista AF, Xavier DJ et al (2013a) Identifying common and specific microRNAs expressed in peripheral blood mononuclear cell of type 1, type 2, and gestational diabetes mellitus patients. BMC Res Notes 6:491

Collares CV, Evangelista AF, Xavier DJ et al (2013b) Transcriptome meta-analysis of peripheral lymphomononuclear cells indicates that gestational diabetes is closer to type 1 diabetes than to type 2 diabetes mellitus. Mol Biol Rep 40:5351–5358

Colli ML, Moore F, Gurzov EN et al (2010) MDA5 and PTPN2, two candidate genes for type 1 diabetes, modify pancreatic b-cell responses to the viral by-product double-stranded RNA. Hum Mol Genet 19:135–146

Concannon P, Rich SS, Nepom GT (2009) Genetics of type 1A diabetes. N Engl J Med 360:1646–1654

Costa V, Aprile M, Esposito R et al (2013) RNA-seq and human complex diseases: recent accomplishments and future perspectives. Eur J Hum Genet 21:134–142

Cox MB, Cairns MJ, Gandhi KS et al (2010) MicroRNAs miR-17 and miR-20a inhibit T cell activation genes and are under-expressed in MS whole blood. PLoS One 5:e12132

Crow MK (2014) Advances in understanding the role of type I interferons in systemic lupus erythematosus. Curr Opin Rheumatol 26:467–474

Crow MK, Kirou KA (2004) Interferon-alpha in systemic lupus erythematosus. Curr Opin Rheumatol 16:541–547

Crow MK, Wohlgemuth J (2003) Microarray analysis of gene expression in lupus. Arthritis Res Ther 5:279–287

Dai Y, Huang YS, Tang M et al (2007) Microarray analysis of microRNA expression in peripheral blood cells of systemic lupus erythematosus patients. Lupus 16:939–946

De Santis G, Ferracin M, Biondani A et al (2010) Altered miRNA expression in T regulatory cells in course of multiple sclerosis. J Neuroimmunol 226:165–171

Deapen D, Escalante A, Weinrib L et al (1992) A revised estimate of twin concordance in systemic lupus erythematosus. Arthritis Rheum 35:311–318

Deng Y, Tsao BP (2010) Genetic susceptibility to systemic lupus erythematosus in the genomic era. Nat Rev Rheumatol 6:683–692

Deng C, Kaplan MJ, Yang J et al (2001) Decreased Ras-mitogen-activated protein kinase signaling may cause DNA hypomethylation in T lymphocytes from lupus patients. Arthritis Rheum 44:397–407

Diaz-Valencia PA, Bougnères P, Valleron AJ (2015) Global epidemiology of type 1 diabetes in young adults and adults: a systematic review. BMC Public Health 15:255

DiMeglio LA, Evans-Molina C, Oram RA (2018) Type 1 diabetes. Lancet 391:2449–2462

Ding S, Liang Y, Zhao M et al (2012) Decreased microRNA-142-3p/5p expression causes CD4+ T cell activation and B cell hyperstimulation in systemic lupus erythematosus. Arthritis Rheum 64:2953–2963

Dobson R, Giovannoni G (2019) Multiple sclerosis – a review. Eur J Neurol 26:27–40

Doria A, Canova M, Tonon M et al (2008) Infections as triggers and complications of systemic lupus erythematosus. Autoimmun Rev 8:24–28

Dorrell C, Schug J, Lin CF et al (2011) Transcriptomes of the major human pancreatic cell types. Diabetologia 54:2832–2844

Du C, Liu C, Kang J et al (2009) MicroRNA miR-326 regulates TH-17 differentiation and is associated with the pathogenesis of multiple sclerosis. Nat Immunol 10:1252–1259

Dumortier O, Van Obberghen E (2012) MicroRNAs in pancreas development. Diabetes Obes Metab 14(Suppl 3):22–28

Dyment DA, Ebers GC, Sadovnick AD (2004) Genetics of multiple sclerosis. Lancet Neurol 3:104–110

Edwards M, Dai R, Ahmed SA (2018) Our environment shapes us: the importance of environment and sex differences in regulation of autoantibody production. Front Immunol 9:478

Eizirik DL, Colli ML, Ortis F (2009) The role of inflammation in insulitis and b-cell loss in type 1 diabetes. Nat Rev Endocrinol 5:219–226

Eizirik DL, Sammeth M, Bouckenooghe T et al (2012) The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. PLoS Genet 8:e1002552

El Ouaamari A, Baroukh N, Martens GA et al (2008) MiR-375 targets 3′-phosphoinositide-dependent protein kinase-1 and regulates glucose-induced biological responses in pancreatic beta-cells. Diabetes 57:2708–2017

Epplen C, Jackel S, Santos EJ et al (1997) Genetic predisposition to multiple sclerosis as revealed by immunoprinting. Ann Neurol 41:341–352

Erener S, Mojibian M, Fox JK et al (2013) Circulating miR-375 as a biomarker of β-cell death and diabetes in mice. Endocrinology 154:603–608

Evangelista AF, Collares CV, Xavier DJ et al (2014) Integrative analysis of the transcriptome profiles observed in type 1, type 2 and gestational diabetes mellitus reveals the role of inflammation. BMC Med Genet 23(7):28. https://doi.org/10.1186/1755-8794-7-28. PMID: 24885568; PMCID: PMC4066312

Fan W, Liang D, Tang Y et al (2012) Identification of microRNA-31 as a novel regulator contributing to impaired interleukin-2 production in T cells from patients with systemic lupus erythematosus. Arthritis Rheum 64:3715–3725

Feng X, Wu H, Grossman JM et al (2006) Association of increased interferon-inducible gene expression with disease activity and lupus nephritis in patients with systemic lupus erythematosus. Arthritis Rheum 54:2951–2962

Fenoglio C, Ridolfi E, Cantoni C et al (2013) Decreased circulating miRNA levels in patients with primary progressive multiple sclerosis. Mult Scler 19:1938–1942

Fornari TA, Donate PB, Macedo C et al (2011) Development of type 1 diabetes mellitus in non-obese diabetic mice follows changes in thymocyte and peripheral T lymphocyte transcriptional activity. Clin Dev Immunol 2011:158735

Foster MH, Kelley VR (1999) Lupus nephritis: update on pathogenesis and disease mechanisms. Semin Nephrol 19:173–181

Frost RJ, Olson EN (2011) Control of glucose homeostasis and insulin sensitivity by the let-7 family of microRNAs. Proc Natl Acad Sci U S A 108:21075–21080

Frota ER, Rodrigues DH, Donadi EA et al (2009) Increased plasma levels of brain derived neurotrophic factor (BDNF) after multiple sclerosis relapse. Neurosci Lett 460:130–132

Furukawa H, Oka S, Matsui T et al (2013) Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus. Hum Immunol 74:170–175

Gandhi KS, McKay FC, Cox M et al (2010) The multiple sclerosis whole blood mRNA transcriptome and genetic associations indicate dysregulation of specific T cell pathways in pathogenesis. Hum Mol Genet 19:2134–2143

Gao X, Zhao S (2020) miRNA-16-5p inhibits the apoptosis of high glucose-induced pancreatic β cells via targeting of CXCL10: potential biomarkers in type 1 diabetes mellitus. Endokrynol Pol 71:404–410

Geenen V, Mottet M, Dardenne O et al (2010) Thymic self-antigens for the design of a negative/tolerogenic self-vaccination against type 1 diabetes. Curr Opin Pharmacol 10:461–472

Ghodke-Puranik Y, Niewold TB (2013) Genetics of the type I interferon pathway in systemic lupus erythematosus. Int J Clin Rheumtol 8. https://doi.org/10.2217/ijr.13.58

Ghodke-Puranik Y, Niewold TB (2015) Immunogenetics of systemic lupus erythematosus: a comprehensive review. J Autoimmun 64:125–136. https://doi.org/10.1016/j.jaut.2015.08.004. Epub 2015 Aug 29. PMID: 26324017; PMCID: PMC4628859

Gough SC, Simmonds MJ (2007) The HLA region and autoimmune disease: associations and mechanisms of action. Curr Genomics 8:453–465

Graham RR, Hom G, Ortmann W et al (2009) Review of recent genome-wide association scans in lupus. J Intern Med 265:680–688

Grammatikos AP, Kyttaris VC, Kis-Toth K et al (2014) A T cell gene expression panel for the diagnosis and monitoring of disease activity in patients with systemic lupus erythematosus. Clin Immunol 150:192–200

Grinberg-Bleyer Y, Baeyens A, You S et al (2010) IL-2 reverses established type 1 diabetes in NOD mice by a local effect on pancreatic regulatory T cells. J Exp Med 207:1871–1878

Guay C, Kruit JK, Rome S et al (2019) Lymphocyte-derived exosomal microRNAs promote pancreatic β cell death and may contribute to type 1 diabetes development. Cell Metab 29:348–361.e6

Guerra SG, Vyse TJ, Cunninghame Graham DS (2012) The genetics of lupus: a functional perspective. Arthritis Res Ther 14:211

Guo J, Casolaro V, Seto E et al (2001) Yin-Yang 1 activates interleukin-4 gene expression in T cells. J Biol Chem 276:48871–48878

Guo J, Lin X, Williams MA et al (2008) Yin-Yang 1 regulates effector cytokine gene expression and T(H)2 immune responses. J Allergy Clin Immunol 122:195–201

Gurevich M, Achiron A (2012) The switch between relapse and remission in multiple sclerosis: continuous inflammatory response balanced by Th1 suppression and neurotrophic factors. J Neuroimmunol 252:83–88

Han GM, Chen SL, Shen N et al (2003) Analysis of gene expression profiles in human systemic lupus erythematosus using oligonucleotide microarray. Genes Immun 4:177–186

Han D, Leyva CA, Matheson D et al (2011) Immune profiling by multiple gene expression analysis in patients at-risk and with type 1 diabetes. Clin Immunol 139:290–301

Harirchian MH, Fatehi F, Sarraf P et al (2018) Worldwide prevalence of familial multiple sclerosis: a systematic review and meta-analysis. Multi Scler Relat Disord 20:43–47

Hecker M, Thamilarasan M, Koczan D et al (2013) MicroRNA expression changes during interferon-beta treatment in the peripheral blood of multiple sclerosis patients. Int J Mol Sci 14:16087–16110

Hennessy E, Clynes M, Jeppesen PB et al (2010) Identification of microRNAs with a role in glucose stimulated insulin secretion by expression profiling of MIN6 cells. Biochem Biophys Res Commun 396:457–462

Herold KC, Brooks-Worrell B, Palmer J et al (2009) Validity and reproducibility of measurement of islet autoreactivity by T-cell assays in subjects with early type 1 diabetes. Diabetes 58:2588–2595

Hollenbach JA, Oksenberg JR (2015) The immunogenetics of multiple sclerosis: a comprehensive review. J Autoimmun 64:13–25

Holler CJ, Webb RL, Laux AL et al (2012) BACE2 expression increases in human neurodegenerative disease. Am J Pathol 180:337–350

Honardoost MA, Kiani-Esfahani A, Ghaedi K et al (2014) miR-326 and miR-26a, two potential markers for diagnosis of relapse and remission phases in patient with relapsing-remitting multiple sclerosis. Gene 544:128–133

Islam T, Gauderman WJ, Cozen W et al (2006) Differential twin concordance for multiple sclerosis by latitude of birthplace. Ann Neurol 60:56–64

Jarvinen P, Kaprio J, Makitalo R et al (1992) Systemic lupus erythematosus and related systemic diseases in a nationwide twin cohort: an increased prevalence of disease in MZ twins and concordance of disease features. J Intern Med 231:67–72

Jayaraman S, Patel A, Jayaraman A et al (2013) Transcriptome analysis of epigenetically modulated genome indicates signature genes in manifestation of type 1 diabetes and its prevention in NOD mice. PLoS One 8:e55074

Jernas M, Malmeström C, Axelsson M et al (2013) MS risk genes are transcriptionally regulated in CSF leukocytes at relapse. Mult Scler 19:403–410

Junker A, Krumbholz M, Eisele S et al (2009) MicroRNA profiling of multiple sclerosis lesions identifies modulators of the regulatory protein CD47. Brain 132:3342–3352

Kaimen-Maciel DR, Reiche EM, Borelli SD et al (2009) HLA-DRB1* allele-associated genetic susceptibility and protection against multiple sclerosis in Brazilian patients. Mol Med Rep 2:993–998

Kamradt T, Mitchison NA (2001) Tolerance and autoimmunity. N Engl J Med 344:655–664

Karpuj MV, Steinman L, Oksenberg JR (1997) Multiple sclerosis: a polygenic disease involving epistatic interactions, germline rearrangements and environmental effects. Neurogenetics 1:21–28

Karumuthil-Melethil S, Perez N, Li R et al (2008) Induction of innate immune response through TLR2 and dectin 1 prevents type 1 diabetes. J Immunol 181:8323–8334

Kaul A, Gordon C, Crow MK et al (2016) Systemic lupus erythematosus. Nat Rev Dis Primers 2:16039

Kawai T, Akira S (2006) TLR signaling. Cell Death Differ 13:816–825

Keller A, Leidinger P, Lange J et al (2009) Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing remitting disease from healthy controls. PLoS One 4:e7440

Keller A, Leidinger P, Steinmeyer F et al (2014) Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. Mult Scler 20:295–303

Kiefer K, Oropallo MA, Cancro MP et al (2012) Role of type I interferons in the activation of autoreactive B cells. Immunol Cell Biol 90:498–504

Kilpinen H, Dermitzakis ET (2012) Genetic and epigenetic contribution. Hum Mol Genet 21:R24–R28

Kirou KA, Lee C, George S et al (2004) Coordinate overexpression of interferon-alpha-induced genes in systemic lupus erythematosus. Arthritis Rheum 50:3958–3967

Kirou KA, Lee C, George S et al (2005) Activation of the interferon-alpha pathway identifies a subgroup of systemic lupus erythematosus patients with distinct serologic features and active disease. Arthritis Rheum 52:1491–1503

Klein SL, Flanagan KL (2016) Sex differences in immune responses. Nat Rev Immunol 16:626–638

Knip M, Siljander H (2008) Autoimmune mechanisms in type 1 diabetes. Autoimmun Rev 7:550–557

Koga M, Kawasaki A, Ito I et al (2011) Cumulative association of eight susceptibility genes with systemic lupus erythematosus in a Japanese female population. J Hum Genet 2011:12

Krützfeldt J, Stoffel M (2006) MicroRNAs: a new class of regulatory genes affecting metabolism. Cell Metab 4:9–12

Kuo CC, Lin SC (2007) Altered FOXO1 transcript levels in peripheral blood mononuclear cells of systemic lupus erythematosus and rheumatoid arthritis patients. Mol Med 13:561–566

Kutlu B, Burdick D, Baxter D et al (2009) Detailed transcriptome atlas of the pancreatic beta cell. BMC Med Genet 2:3

Langefeld CD, Ainsworth HC, Cunninghame Graham DS et al (2017) Transancestral mapping and genetic load in systemic lupus erythematosus. Nat Commun 8:16021

Laplaud DA, Ruiz C, Wiertlewski S et al (2004) Blood T-cell receptor beta chain transcriptome in multiple sclerosis. Characterization of the T cells with altered CDR3 length distribution. Brain 127:981–995

Lashine YA, Seoudi AM, Salah S et al (2011) Expression signature of microRNA-181-a reveals its crucial role in the pathogenesis of paediatric systemic lupus erythematosus. Clin Exp Rheumatol 29:351–357

Latreille M, Hausser J, Stützer I et al (2014) MicroRNA-7a regulates pancreatic β cell function. J Clin Invest 124:2722–2735

Layer K, Lin G, Nencioni A et al (2003) Autoimmunity as the consequence of a spontaneous mutation in Rasgrp1. Immunity 19:243–255

Lee HM, Sugino H, Aoki C et al (2011) Underexpression of mitochondrial-DNA encoded ATP synthesis-related genes and DNA repair genes in systemic lupus erythematosus. Arthritis Res Ther 13:R63

Lehmann-Werman R, Neiman D, Zemmour H, Moss J et al (2016) Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci U S A 113:E1826–E1834

Lei J, Li J, Li X et al (2020) CT imaging of the 2019 novel coronavirus (2019-nCoV) pneumonia. Radiology 295(1):18

Li R, Perez N, Karumuthil-Melethil S et al (2007) Bone marrow is a preferential homing site for autoreactive T-cells in type 1 diabetes. Diabetes 56:2251–2259

Li QZ, Zhou J, Lian Y et al (2010) Interferon signature gene expression is correlated with autoantibody profiles in patients with incomplete lupus syndromes. Clin Exp Immunol 159:281–291

Lin T, Zhang D, Liu X et al (2016) Parental care improves immunity in the seahorse (Hippocampus erectus). Fish Shellfish Immunol 58:554–562

Lindberg RL, Hoffmann F, Mehling M et al (2010) Altered expression of miR-17-5p in CD4+ lymphocytes of relapsing– remitting multiple sclerosis patients. Eur J Immunol 40:888–898

Lindén M, Khademi M, Lima Bomfim I et al (2013) Multiple sclerosis risk genotypes correlate with an elevated cerebrospinal fluid level of the suggested prognostic marker CXCL13. Mult Scler 19:863–870

Lit LC, Wong CK, Li EK et al (2007) Elevated gene expression of Th1/Th2 associated transcription factors is correlated with disease activity in patients with systemic lupus erythematosus. J Rheumatol 34:89–96

Liu Z, Davidson A (2012) Taming lupus – a new understanding of pathogenesis is leading to clinical advances. Nat Med 18:871–882

Lood C, Amisten S, Gullstrand B et al (2010) Platelet transcriptional profile and protein expression in patients with systemic lupus erythematosus: up-regulation of the type I interferon system is strongly associated with vascular disease. Blood 116:1951–1957

Lu MC, Lai NS, Chen HC et al (2013) Decreased microRNA(miR)-145 and increased miR-224 expression in T cells from patients with systemic lupus erythematosus involved in lupus immunopathogenesis. Clin Exp Immunol 171:91–99

Lublin FD, Reingold SC (1996) Defining the clinical course of multiple sclerosis: results of an international survey. Neurology 46:907–911

Lucchinetti CF, Brueck W, Rodriguez M et al (1998) Multiple sclerosis: lessons from neuropathology. Semin Neurol 18:337–349

Lyons PA, McKinney EF, Rayner TF et al (2010) Novel expression signatures identified by transcriptional analysis of separated leukocyte subsets in systemic lupus erythematosus and vasculitis. Ann Rheum Dis 69:1208–1213

Maahs DM, West NA, Lawrence JM et al (2010) Epidemiology of type 1 diabetes. Endocrinol Metab Clin N Am 39:481–497

Maas K, Chen H, Shyr Y et al (2005) Shared gene expression profiles in individuals with autoimmune disease and unaffected first-degree relatives of individuals with autoimmune disease. Hum Mol Genet 14:1305–1314

Mackay IR (2009) Clustering and commonalities among autoimmune diseases. J Autoimmun 33:170–177

Mandolesi G, De Vito F, Musella A et al (2017) miR-142-3p Is a Key regulator of IL-1β-dependent synaptopathy in neuroinflammation. J Neurosci 37:546–561

Massaro JD, Polli CD, Silva MCE et al (2019) Post-transcriptional markers associated with clinical complications in Type 1 and Type 2 diabetes mellitus. Mol Cell Endocrinol 490:1–14

Mattick JS, Makunin IV (2006) Noncoding RNA. Hum Mol Genet 15(Spec 1):R17–R29

Matzinger P (2002) The danger model: a renewed sense of self. Science 296:301–305

Mayer-Davis EJ, Lawrence JM, Dabelea D et al (2017) Incidence trends of type 1 and type 2 diabetes among youths, 2002–2012. N Engl J Med 376:1419–1429

McCarthy EM, Smith S, Lee RZ et al (2014) The association of cytokines with disease activity and damage scores in systemic lupus erythematosus patients. Rheumatology (Oxford) 53:1586–1594

McDevitt HO, Unanue ER (2008) Autoimmune diabetes mellitus–much progress, but many challenges. Adv Immunol 100:1–12

McLaughlin KA, Richardson CC, Ravishankar A et al (2016) Identification of tetraspanin-7 as a target of autoantibodies in type 1 diabetes. Diabetes 65:1690–1698

Melief J, de Wit SJ, Van Eden CG et al (2013) HPA axis activity in multiple sclerosis correlates with disease severity, lesion type and gene expression in normal-appearing white matter. Acta Neuropathol 126:237–249

Melzer S, Michael M, Caputi A et al (2012) Long-range-projecting GABAergic neurons modulate inhibition in hippocampus and entorhinal cortex. Science 335:1506–1510

Mokry LE, Ross S, Timpson NJ et al (2016) Obesity and multiple sclerosis: a Mendelian randomization study. PLoS Med 13:e1002053

Moore F, Colli ML, Cnop M et al (2009) PTPN2, a candidate gene for type 1 diabetes, modulates interferon-c-induced pancreatic b-cell apoptosis. Diabetes 58:1283–1291

Moore CS, Rao VT, Durafourt BA et al (2013) miR-155 as a multiple sclerosis-relevant regulator of myeloid cell polarization. Ann Neurol 74:709–720

Morris DL, Fernando MM, Taylor KE et al (2014) MHC associations with clinical and autoantibody manifestations in European SLE. Genes Immun 15:210–217

Morris DL, Sheng Y, Zhang Y et al (2016) Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. Nat Genet 48:940–946

Moser KL, Kelly JA, Lessard CJ et al (2009) Recent insights into the genetic basis of systemic lupus erythematosus. Genes Immun 10:373–379

Moulton VR (2018) Sex hormones in acquired immunity and autoimmune disease. Front Immunol 9:2279

Mumford CJ, Wood NW, Kellar-Wood H et al (1994) The British Isles survey of multiple sclerosis in twins. Neurology 44:11–15

Mutz K-O, Heilkenbrinker A, Lonne M et al (2013) Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol 24:22–30

Nakou M, Knowlton N, Frank MB et al (2008) Gene expression in systemic lupus erythematosus: bone marrow analysis differentiates active from inactive disease and reveals apoptosis and granulopoiesis signatures. Arthritis Rheum 58(11):3541–3549

Nakou M, Bertsias G, Stagakis I et al (2010) Gene network analysis of bone marrow mononuclear cells reveals activation of multiple kinase pathways in human systemic lupus erythematosus. PLoS One 5:e13351

Nerup J, Nierras C, Plagnol V et al (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703–707

Nielsen NM, Westergaard T, Rostgaard K et al (2005) Familial risk of multiple sclerosis: a nationwide cohort study. Am J Epidemiol 162:774–778

Niewold TB (2011) Interferon alpha as a primary pathogenic factor in human lupus. J Interf Cytokine Res 31:887–892

Niewold TB (2015) Advances in lupus genetics. Curr Opin Rheumatol 27:440–447

Niewold TB, Swedler WI (2005) Systemic lupus erythematosus arising during interferon-alpha therapy for cryoglobulinemic vasculitis associated with hepatitis C. Clin Rheumatol 24:178–181

Niewold TB, Hua J, Lehman TJ et al (2007) High serum IFN-alpha activity is a heritable risk factor for systemic lupus erythematosus. Genes Immun 8:492–502

Noble JA (2015) Immunogenetics of type 1 diabetes: a comprehensive review. J Autoimmun 64:101–112

Noorbakhsh F, Ellestad KK, Maingat F et al (2011) Impaired neurosteroid synthesis in multiple sclerosis. Brain 134:2703–2721

Noseworthy JH, Lucchinetti C, Rodriguez M et al (2000) Multiple sclerosis. N Engl J Med 343:938–952

Novak J, Lehuen A (2011) Mechanism of regulation of autoimmunity by iNKT cells. Cytokine 53:263–270

Oh J, Broyles SS (2005) Host cell nuclear proteins are recruited to cytoplasmic vaccinia virus replication complexes. J Virol 79:12852–12860

Oksenberg JR, Baranzini SE, Sawcer S et al (2008) The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. Nat Rev Genet 9:516–526

Orton SM, Herrera BM, Yee IM et al (2006) Sex ratio of multiple sclerosis in Canada: a longitudinal study. Lancet Neurol 5:932–936

Ortona E, Pierdominici M, Maselli A et al (2016) Sex-based differences in autoimmune diseases. Ann Ist Super Sanita 52:205–212

Pan W, Zhu S, Yuan M et al (2010) MicroRNA-21 and microRNA-148a contribute to DNA hypomethylation in lupus CD4+ T cells by directly and indirectly targeting DNA methyltransferase 1. J Immunol 184:6773–6781

Panousis NI, Bertsias GK, Ongen H et al (2019) Combined genetic and transcriptome analysis of patients with SLE: distinct, targetable signatures for susceptibility and severity. Ann Rheum Dis 78:1079–1089

Peeters LM, Parciak T, Walton C et al (2020) COVID-19 in people with multiple sclerosis: a global data sharing initiative. Mult Scler 26:1157–1162

Plagnol V, Howson JM, Smyth DJ et al (2011) Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet 7:e1002216

Plaisance V, Abderrahmani A, Perret-Menoud V et al (2006) MicroRNA-9 controls the expression of granuphilin/Slp4 and the secretory response of insulin-producing cells. J Biol Chem 281:26932–26942

Planas R, Pujol-Borrell R, Vives-Pi M (2010) Global gene expression changes in type 1 diabetes: insights into autoimmune response in the target organ and in the periphery. Immunol Lett 133:55–61

Pociot F, Akolkar B, Concannon P et al (2010) Genetics of type 1 diabetes: what's next? Diabetes 59:1561–1571

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–641

Poo MM (2001) Neurotrophins as synaptic modulators. Nat Rev Neurosci 2:24–32

Poy MN, Eliasson L, Krutzfeldt J et al (2004) A pancreatic islet-specific microRNA regulates insulin secretion. Nature 432:226–230

Pugliese A, Miceli D (2002) The insulin gene in diabetes. Diabetes Metab Res Rev 18:13–25

Purohit S, She JX (2008) Biomarkers for type 1 diabetes. Int J Clin Exp Med 1:98–116

Qin H, Zhu X, Liang J et al (2013) MicroRNA-29b contributes to DNA hypomethylation of CD4+ T cells in systemic lupus erythematosus by indirectly targeting DNA methyltransferase 1. J Dermatol Sci 69:61–67

Ramagopalan SV, Dobson R, Meier UC et al (2010) Multiple sclerosis: risk factors, prodromes, and potential causal pathways. Lancet Neurol 9:727–739

Rassi DM, Junta CM, Fachin AL et al (2008) Gene expression profiles stratified according to type 1 diabetes mellitus susceptibility regions. Ann N Y Acad Sci 1150:282–289

Rees F, Doherty M, Grainge MJ et al (2017) The worldwide incidence and prevalence of systemic lupus erythematosus: a systematic review of epidemiological studies. Rheumatology (Oxford) 56:1945–1961

Remoli ME, Ragimbeau J, Giacomini E et al (2007) NF-{kappa}B is required for STAT-4 expression during dendritic cell maturation. J Leukoc Biol 81:355–363

Reynier F, Pachot A, Paye M et al (2010) Specific gene expression signature associated with development of autoimmune type-I diabetes using whole-blood microarray analysis. Genes Immun 11:269–278

Ridolfi E, Fenoglio C, Cantoni C et al (2013) Expression and genetic analysis of MicroRNAs Involved in Multiple Sclerosis. Int J Mol Sci 14:4375–4384

Riveros C, Mellor D, Gandhi KS et al (2010) A transcription factor map as revealed by a genome-wide gene expression analysis of whole-blood mRNA transcriptome in multiple sclerosis. PLoS One 5:e14176

Ronnblom LE, Alm GV, Oberg KE (1990) Possible induction of systemic lupus erythematosus by interferon-alpha treatment in a patient with a malignant carcinoid tumour. J Intern Med 227:207–210

Rose NR (2016) Prediction and prevention of autoimmune disease in the 21st century: a review and preview. Am J Epidemiol 183:403–406

Rus V, Chen H, Zernetkina V et al (2004) Gene expression profiling in peripheral blood mononuclear cells from lupus patients with active and inactive disease. Clin Immunol 112:231–234

Sadovnick AD, Ebers GC (1993) Epidemiology of multiple sclerosis: a critical overview. Can J Neurol Sci 20:17–29

Sadovnick AD, Ebers GC, Dyment DA et al (1996) Evidence for genetic basis of multiple sclerosis. Lancet 347:1728–1730

Sandrin-Garcia P, Brandão LA, Guimarães RL et al (2012) Functional single-nucleotide polymorphisms in the DEFB1 gene are associated with systemic lupus erythematosus in Southern Brazilians. Lupus 21:625–631

Santiago-Raber ML, Lawson BR, Dummer W et al (2001) Role of cyclin kinase inhibitor p21 in systemic autoimmunity. J Immunol 167:4067–4074

Satoh J, Misawa T, Tabunoki H et al (2008) Molecular network analysis of T-cell transcriptome suggests aberrant regulation of gene expression by NF-kappaB as a biomarker for relapse of multiple sclerosis. Dis Markers 25:27–35

Schwartzman-Morris J, Putterman C (2012) Gender differences in the pathogenesis and outcome of lupus and of lupus nephritis. Clin Dev Immunol 2012:604892

Shaikh MF, Jordan N, D’Cruz DP (2017) Systemic lupus erythematosus. Clin Med (Lond) 17:78–83

Shaw JE, Sicree RA, Zimmet PZ (2010) Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Res Clin Pract 87:4–14

Sia C (2006) Replenishing peripheral CD4(+) regulatory T cells: a possible immune-intervention strategy in type 1 diabetes? Rev Diabet Stud 3:102–107

Sintzel MB, Rametta M, Reder AT (2018) Vitamin D and multiple sclerosis: a comprehensive review. Neurol Ther 7:59–85

Sondergaard HB, Hesse D, Krakauer M et al (2013) Differential microRNA expression in blood in multiple sclerosis. Mult Scler 19:1849–1857

Stagakis E, Bertsias G, Verginis P et al (2011) Identification of novel microRNA signatures linked to human lupus disease activity and pathogenesis: miR-21 regulates aberrant T cell responses through regulation of PDCD4 expression. Ann Rheum Dis 70:1496–1506

Stone RC, Du P, Feng D et al (2013) RNA-Seq for enrichment and analysis of IRF5 transcript expression in SLE. PLoS One 8:e54487

Straub RH (2007) The complex role of estrogens in inflammation. Endocr Rev 28:521–574

Sui WG, Lin H, Chen JJ et al (2012) Comprehensive analysis of transcription factor expression patterns in peripheral blood mononuclear cell of systemic lupus erythematosus. Int J Rheum Dis 15:212–219

Tajouri L, Fernandez F, Griffiths LR (2007) Gene expression studies in multiple sclerosis. Curr Genomics 8:181–189

Tang X, Muniappan L, Tang G et al (2009a) Identification of glucose-regulated miRNAs from pancreatic {beta} cells reveals a role for miR-30d in insulin transcription. RNA 15:287–293

Tang Y, Luo X, Cui H et al (2009b) MicroRNA-146A contributes to abnormal activation of the type I interferon pathway in human lupus by targeting the key signaling proteins. Arthritis Rheum 60:1065–1075

Tang Q, Yang Y, Zhao M et al (2015) Mycophenolic acid upregulates miR-142-3P/5P and miR-146a in lupus CD4+T cells. Lupus 24:935–942

Teruel R, Corral J, Pérez-Andreu V et al (2011) Potential role of miRNAs in developmental haemostasis. PLoS One 6:e17648

Tisch R, Wang B (2008) Dysrulation of T cell peripheral tolerance in type 1 diabetes. Adv Immunol 100:125–149

Todd JA (2010) Etiology of type 1 diabetes. Immunity 32:457–467

Toukap AN, Galant C, Theate I et al (2007) Identification of distinct gene expression profiles in the synovium of patients with systemic lupus erythematosus. Arthritis Rheum 56:1579–1588

Trombetta AC, Meroni M, Cutolo M (2017) Steroids and autoimmunity. Front Horm Res 48:121–132

Voulgarelis M, Giannouli S, Tasidou A et al (2006) Bone marrow histological findings in systemic lupus erythematosus with hematological abnormalities: a clinicopathological study. Am J Hematol 81:590–597

Wang G, Tam LS, Li EK et al (2010) Serum and urinary cell-free MiR-146a and MiR-155 in patients with systemic lupus erythematosus. J Rheumatol 37:2516–2522

Wang G, Tam LS, Li EK et al (2011) Serum and urinary free microRNA level in patients with systemic lupus erythematosus. Lupus 20:493–500

Wang H, Peng W, Ouyang X et al (2012) Circulating microRNAs as candidate biomarkers in patients with systemic lupus erythematosus. Transl Res 160:198–206

Webber D, Cao J, Dominguez D et al (2020) Association of systemic lupus erythematosus (SLE) genetic susceptibility loci with lupus nephritis in childhood-onset and adult-onset SLE. Rheumatology (Oxford) 59:90–98

Weckerle CE, Niewold TB (2011) The unexplained female predominance of systemic lupus erythematosus: clues from genetic and cytokine studies. Clin Rev Allergy Immunol 40:42–49

Weckerle CE, Franek BS, Kelly JA et al (2011) Network analysis of associations between serum interferon-alpha activity, autoantibodies, and clinical features in systemic lupus erythematosus. Arthritis Rheum 63:1044–1053

Weinshenker BG (1994) Natural history of multiple sclerosis. Ann Neurol 36:S6–S11

Wenzlau JM, Juhl K, Yu L et al (2007) The cation efflux transporter ZnT8 (Slc30A8) is a major autoantigen in human type 1 diabetes. Proc Natl Acad Sci U S A 104:17040–17045

Willer CJ, Dyment DA, Risch NJ et al (2003) Twin concordance and sibling recurrence rates in multiple sclerosis. Proc Natl Acad Sci U S A 100:12877–12882

Yao Y, Higgs BW, Morehouse C et al (2009) Development of potential pharmacodynamic and diagnostic markers for anti-IFN-α monoclonal antibody trials in systemic lupus erythematosus. Hum Genomics Proteomics pii: 374312

Yuan Y, Kasar S, Underbayev C et al (2012) Role of microRNA-15a in autoantibody production in interferon-augmented murine model of lupus. Mol Immunol 52:61–70

Zhao X, Tang Y, Qu B et al (2010) MicroRNA-125a contributes to elevated inflammatory chemokine RANTES levels via targeting KLF13 in systemic lupus erythematosus. Arthritis Rheum 62:3425–3435

Zhao S, Wang Y, Liang Y et al (2011) MicroRNA-126 regulates DNA methylation in CD4+ T cells and contributes to systemic lupus erythematosus by targeting DNA methyl-transferase 1. Arthritis Rheum 63:1376–1386

Zhernakova A, van Diemen CC, Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet 10:43–55

Ziegler AG, Rewers M, Simell O et al (2013) Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. JAMA 309:2473–2479

Chapter 12
Transcriptome Profiling in Experimental Inflammatory Arthritis

Olga Martinez Ibañez, José Ricardo Jensen, and Marcelo De Franco

12.1 Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disease that affects 0.5 to 1% of the human population. RA is a complex pathology characterized by systemic chronic inflammation with the accumulation into synovium and periarticular spaces of activated T and B lymphocytes, innate immune cells such as neutrophils, mast cells, dendritic cells, natural killer cells, and macrophages, and endothelial cells. Rheumatoid fibroblast-like synoviocytes, which exhibit invasive characteristics and synovial macrophages with proinflammatory properties are crucial for the progression of arthritis causing proliferation of synovial membranes and the formation of the invasive pannus that erodes cartilage and bone. In human patients the clinical signs of RA are largely heterogeneous, but the disease is considered to be autoimmune (You et al. 2014). RA heterogeneity is demonstrated by the presence of distinct autoantibody specificities, such as antibodies against immunoglobulins, the rheumatoid factor (RF), and anti-cyclic citrullinated peptide antibodies (ACPA) in the serum, the differential responsiveness to treatment, and by the variability in clinical signs (Silman and Pearson 2002). The precise etiology of RA remains poorly understood, but the main symptoms are chronic synovitis, joint erosion, and several immune abnormalities in both the innate and adaptive compartments.

Given the complexity of RA, systems biology approaches designed to give a general view of different aspects of the disease are required to better understand the basis of arthritis. Oligonucleotide-based microarray technology for global gene

O. M. Ibañez · J. R. Jensen · M. De Franco (✉)
Laboratory of Immunogenetics, Butantan Institute, São Paulo, São Paulo, Brazil

Diagnostic section, Pasteur Institute, São Paulo, São Paulo, Brazil
e-mail: mdfranco@pasteur.saude.sp.gov.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_12

expression profiling has arisen as a powerful tool to investigate the molecular complexity and pathogenesis of arthritis and other complex pathologies. This genomic or transcriptomic method combined with postgenomic techniques provides an opportunity to monitor the complex interactions between genes and environment, the regulation of genes and of RNA transcripts, and the proteins that constitute the basis for the etiology or progression of the diseases (Jarvis and Frank 2010).

Gene expression profiling studies of tissues from RA patients showed marked variation in gene expression profiles that allowed to identify distinct molecular disease mechanisms involved in RA pathology (Baechler et al. 2006). The relative contribution of the different mechanisms may vary among patients and in different stages of the disease. Thus, the broad goals of expression profiling in RA are the improvement of understanding of the pathogenic mechanisms underlying RA, the identification of disease subsets and new drug targets, and the assessment of disease activity, such as responsiveness to therapy, overall disease severity, and organ-specific risk, and development of new diagnostic tests (Teixeira et al. 2009).

Genetic and environmental factors contribute to the development of this disease. Numerous studies have indicated the participation of the major histocompatibility complex (MHC) class II alleles and non-MHC genes, such as the *solute carrier family 11a member 1—SLC11A1* (formerly named *NRAMP1*- Natural resistance-associated macrophage protein 1) related to macrophage activation (Runstadler et al. 2005). Identification of the major roles of the participating cells and of candidate genes has been an important subject of study to the understanding of RA pathogenesis (Kurko et al. 2013).

12.2 Experimental Models of Rheumatoid Arthritis

The initial or preclinical stages of RA are difficult to be studied in humans but numerous arthritis experimental models have been developed which are valuable tools for in-depth investigation of pathogenic pathways that are involved in the several phases of the disease (Kobezda et al. 2014). Regarding ethical procedures, in these models the animals can be submitted to immunizations with arthritogenic substances or antigens, to cell transfer or depletion, to phenotypic selective crosses, to genetic manipulations for the production of transgenic or knockout individuals, etc. Most importantly, these models have been useful for the candidacy of targets for preventive or therapeutic strategies (Asquith et al. 2009).

Several studies have used different animal models for arthritis, generally induced by the injection of adjuvants (AIA), proteoglycan (PGIA), type II collagen (CIA), or pristane (PIA) (Kannan et al. 2005).

Collagen-induced arthritis (CIA). Type II collagen (CII) is expressed exclusively in the articular joint. Although the relationship between anti-CII immunity and human rheumatoid arthritis (RA) has been studied for a long time, definitive conclusions have not been established. CII, as an autoantigen, has been studied extensively in small animal models, such as mice and rats, and the collagen-induced arthritis

(CIA) model has increased our understanding of the pathogenesis of human RA (Cho et al. 2007). The disease is class II MHC restricted but mouse strains with permissive haplotypes vary in their susceptibility to CIA. Arthritis development is associated with B and T lymphocyte responses and the generation of anticollagen antibodies and T-cells.

Collagen antibody-induced arthritis (CAIA) in mice has demonstrated the role of humoral immunity in arthritis development. It has been useful for the identification of collagen epitopes for the generation of arthritogenic antibody cocktails that represent humoral autoimmunity in RA. The disease is characterized by macrophage and polymorphonuclear cell infiltration and no T- and B-cell involvement and is non-MHC class II restricted (Hirose and Tanaka 2011).

Proteoglycan-induced arthritis (PGIA) is based on the immunization of mice with human cartilage-derived proteoglycans, which induces the development of severe polyarthritis and spondylitis (Glant et al. 2003).

Pristane-induced arthritis (PIA) has proven to be a valuable experimental model for inflammatory RA. The natural saturated terpenoid alkane 2,4,6,10-tetramethyl pentadecane induces an acute inflammation followed by a chronic relapsing phase. The reaction is T-cell dependent with edema and articular infiltration of mononuclear and polymorphonuclear cells (Potter and Wax 1981).

There are also genetically manipulated models that develop RA spontaneously. For example, transgenic mice overexpressing human TNF-α develop chronic inflammatory erosive polyarthritis (Li and Schwarz 2003). This model highlights the importance of TNF-α in cytokine network in RA. Another example is the IL1 receptor antagonist-deficient mouse that develops inflammatory arthritis mediated by a polarized TH17 response (van den Berg 2009; Lubberts et al. 2005).

In experimental models, microarray analysis should optimally be carried out in isolated populations of cells. New methods have emerged for transcriptomic analysis that are based on single cell RNA sequencing and high-resolution spatial transcriptomic technology. Although they still have limitations, the methods allow the analysis of the subpopulations of cells that make up the tissue and their location (Reviewed in Carr et al. 2020). However, in complex diseases such as RA there is extensive tissue damage with the contribution of several cell types. Hence the analysis of rodent whole ankle joints or of footpads, which comprise heterogeneous cell types, has given a global view of differential gene expression during the several phases of arthritis onset and development. Differential expression of genes encoding tissue repair factors, signal transduction molecules, transcription factors, and DNA repair enzymes, as well as cell cycle regulators have been observed in multiple microarray experiments. An interesting observation in these experiments is the transcriptome map of the differentially expressed genes; in different models of arthritis there is a functional grouping of dysregulated genes forming clusters in the chromosomes. Examples are the MHC class I and class II gene clusters, known to affect susceptibility to a variety of autoimmune diseases and the chemoattractant gene clusters such as CC or CXC chemokine ligands and receptors, which mediate infiltration of leukocytes into synovial tissue, a hallmark of RA (Fujikado et al. 2006). Some studies attempt to link differentially expressed genes into interactive regulatory networks (Silva et al. 2009). This approach is quite powerful to identify new targets for therapy by looking at the network structures, the places (genes) with the highest connectivity in which disruption would have a larger impact.

12.3 Loci Regulating Inflammatory Arthritis

The identification of the loci influencing inflammatory arthritis in animal models is important for parallel genetic studies in humans. The individual genetic constitution of experimental animals involving major histocompatibility complex (MHC) or non-MHC genes has been associated with variations in rheumatoid arthritis susceptibility. In mice or rats, genome-wide linkage studies with DNA polymorphism markers, such as microsatellites or single nucleotide polymorphisms (SNPs), have been carried out using intercross progenies of resistant and susceptible strains. These studies, in which environmental effects and genetic backgrounds are controlled, have been useful for the study of the genetic basis of RA (Ibrahim and Yu 2006).

Several QTLs (Quantitative Trait Loci) were identified in different models of experimental arthritis. The first locus controlling pristane-induced arthritis (PIA) detected in mice was *Prtia1* on chromosome 3, in an intercross population from mice selected for high and low antibody production (Jensen et al. 2006). QTL was also mapped in other arthritis models such as those induced by *Borrelia burgdorferi* (Roper et al. 2001), PGIA (Glant et al. 2004), and collagen (Adarichev et al. 2003). Nonoverlapping sets of QTLs were identified, generating a heterogeneous picture of risk alleles (Besenyei et al. 2012; Kurko et al. 2013). The results evidence the genetic heterogeneity in the control of the different stages and phenotypes of the disease. Table 12.1 presents some relevant coincident susceptibility QTLs in rheumatoid arthritis, according to GWAS studies carried out in mice and humans.

Numerous RA QTLs have been mapped but few of the associated polymorphisms were identified in protein-coding regions of genes causing changes in protein structure or function. This suggests that polymorphisms in noncoding regions which might affect gene expression largely contribute to variations in RA susceptibility. In this way, transcriptome technology can also be used to detect genetic polymorphisms that regulate gene expression levels.

12.4 Combining Transcriptome and Genome Screening to Identify Genes That Control Arthritis

The two genomic approaches, that is, transcriptome and genome screening (GWAS), have been combined in studies where the locations of differently expressed genes during RA are compared with those mapping at QTLs for arthritis, for immune or inflammatory responses, or for other autoimmune diseases (Yu et al. 2007). The

Table 12.1 Common arthritis-associated QTLs (Non-MHC regions) mapped by GWAS in mice and humans

Mouse			Human	
Chr	Locus name	Candidate gene	Chr position	Locus name
1	*Cia14*	Aff3: expressed in lymphoid cells, encodes a nuclear factor that contains transcriptional activation domains	2q11	AFF3
1	*Cia9,* *Pgia1*	Fcgr2b: a variant allele alters dendritic cell behavior, suggesting a role for dendritic cells in RA pathology	1q23	FCGR2A
2	*Cia2,* *Cia4,* *Pgia2*	Traf1/Hc: Genetic variants associated to risk of anti-CCP antibody-positive RA	9q33	TRAF1/C5
3	*Cia21,* *Cia22,* *Pgia26* *Prtia1*	Cd2: encodes a costimulatory molecule found on natural killer and T cells	1p23	CD2
		Ptpn22: the gene is a negative regulator of T cells. Allele variant affects binding to an intracellular signaling molecule (Csk) resulting in a failure to switch off T cells or to delete auto-reactive T cells during thymic development	1p13	PTPN22
5	*Pgia16,* *Cia13*	Rbjp: The gene encodes a transcription factor involved in the notch signaling pathway and in regulation of T-cell development	4p15	RBJP
6	*Pgia19*	Irf5: transcription factor involved in antiviral and anti-inflammatory responses and in differentiation of B-cells regulation	7q32	IRF5
10	*Pgia6,*	Tnfaip3: knock-out mice develop severe inflammation	6q23	TNFAIP3
	Pgia6b	Prdm1: The gene product is a transcription factor involved in B-cell regulation	6q21	PRDM1
10	*Cia8*	Kif5a: gene encodes a kinesin-heavy chain	12q13	KIF5A
		Pip4k2c: phosphatidyl inositol kinase		PIP4K2C
13	*Pgia15,*	IL6st/Ankrd55: IL-6	5q11	ANKRD55
	Cia19	Signal transduction gene region		IL6ST
15	*Pgia9,* *Cia35,* *Cia37*	IL2rb	22q12	IL2RB
		Bik: apoptosisinducing,BCL2-interacting killer	8p3	BIK
18	*Pgia11*	Ptpn2: KO mice have increased susceptibility to inflammatory diseases	18p11.3-p11.2	PTPN2

Gene names: *Affr* AF4/FMR2 family, member 3; *Fcgr2b* Fc receptor IgG, low affinity IIb; *Traf1* TNF receptor-associated factor 1; *Ptpn22* protein tyrosine phosphatase, nonreceptor type 22 (lymphoid); *Rbjp* recombination signal binding protein for immunoglobulin kappa J region; *Irf5* interferon regulatory factor 5; *Tnfaip3* tumor necrosis factor, alpha-induced protein 3; *Prdm* PR domain containing 1, with ZNF domain; *Kif5a* kinesin family member 5A; *Pip4k2c* phosphatidylinositol-5-phosphate 4-kinase, type II, gamma; *Ankrd55* ankyrin repeat domain 55; *Ptpn2* protein tyrosine phosphatase, nonreceptor type 2 approach has been useful to candidate genes inside the QTLs. The coincidence of chromosomal locations of genes in QTLs in different model systems with the locations of the corresponding human orthologue is a good indicator of their implication in RA control.

Furthermore, the modulation of common genes during RA, irrespective of etiology and of species indicates the importance of these mediators in the pathogenesis of arthritis. For example, the augmented expression of chemokines and receptors, which recruit neutrophils or naïve and memory T cells to inflammatory sites, is very important to disease progression. Chemokines and ligands are found in the synovial tissue of patients with RA; proinflammatory cytokines and their cognate receptors, such as IL-1β, IL-1RI, TNF-α R, IL-6Rα, IL-2Rγ, and IL-17R, are upregulated in several RA models as well as in arthritis patients; IL-1β induces serum amyloid A3 (Saa3) and the matrix metalloproteinases Mmp-3 and Mmp-9 that are also upregulated in several models. High upregulation in runt-related transcription factor 1 (RUNX1) and a group of transporter genes such as solute carrier 11 family A1 (*Slc11a1,* formerly *Nramp1*) is also a common feature in RA models. In synthesis, a remarkable feature that originated from numerous transcriptome or genomic studies of arthritis has been the demonstration of gene expression signatures associated with inflammation. The results evidence that besides being an antigen-driven event there is an important interplay between innate and adaptive immunity systems in the etiology of RA (Jarvis and Frank 2010).

12.5 A Model to Study Inflammatory Rheumatoid Arthritis: AIRmax and AIRmin Phenotypically Selected Mouse Lines

Heterogeneous mice selected for maximal (AIRmax) or minimal (AIRmin) acute inflammatory reaction appeared to be useful models for studying the mechanisms involved in rheumatoid arthritis susceptibility (Vigar et al. 2000).

AIRmax and AIRmin mice were produced by bidirectional selection, starting from a highly polymorphic population (F0) derived from the intercrossing of eight inbred mouse strains (Fig. 12.1). The selection phenotypes chosen were localized leukocyte influx and exudated plasma proteins 24 h after the subcutaneous injection of polyacrylamide beads (Biogel), a nonantigenic, insoluble, and chemically inert substance (Ibanez et al. 1992). The progressive divergence of the AIRmax and AIRmin lines during successive generations of selective breeding reached 20- and 2.5-fold differences in leukocyte infiltration and exudated protein concentrations respectively. These differences resulted from the accumulation of alleles in quantitative trait loci endowed with opposite and additive effects on the inflammatory response. Inbreeding was avoided for selective breeding, and as such AIRmax and AIRmin mice are outbred mice that maintain a heterogeneous genetic background but are homozygous in acute inflammation modifier loci. Analysis of the selective

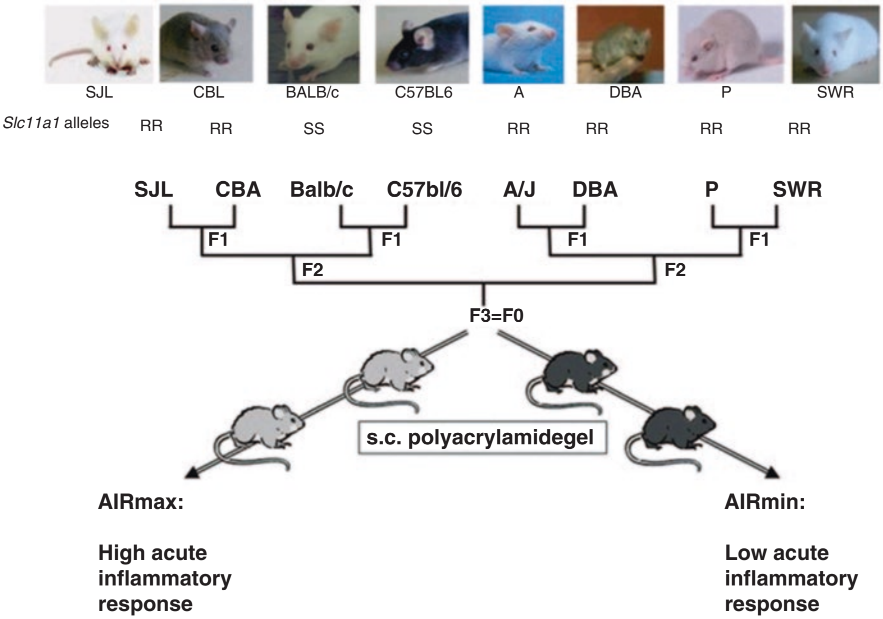


Fig. 12.1 Scheme used for the production of the foundation population (F0) by the intercrossing of eight inbred strains of mice for the production of AIRmax and AIRmin mice by bidirectional phenotypic selection

processes indicated that the AIR phenotype is regulated by at least 11 QTLs (Biozzi et al. 1998).

Pristane-induced arthritis (PIA) has proven to be a valuable experimental model for inflammatory RA for its delayed onset, chronicity, and independence from xenoantigen administration. Thus, arthritis ensues from a sensitization over time and pristane has been described to improve autoimmunity by the activation of the immune response against cross-reactive microbiota antigens (Patten et al. 2004). AIRmax mice are extremely susceptible whereas AIRmin mice are resistant to PIA (Fig. 12.2a). The incidence of PIA in AIRmax mice was similar to that of inbred DBA/1 and BALB/c mice although with higher severity. The incidence and severity were more intense than in the CBA/Igb model because 15 to 25% of these mice develop inflammation of the ankle and wrist joints approximately 200 days after pristane injection. PIA is accompanied by markedly elevated humoral agalactosyl IgG levels mediated by IL6 production (Thompson et al. 1992) and CD4+ T cell (Th)-dependent (Stasiuk et al. 1997) immune responses to mycobacterial 65-kDa heat shock protein (hsp65). Moreover, the protection against PIA is mediated by Th2-associated cytokines produced after hsp65 preimmunization (Thompson et al. 1998; Thompson et al. 1990). In contrast to the immune response profile observed in inbred mice, high IgG1 anti-hsp65 levels were observed in susceptible AIRmax mice, whereas IgG2a was the predominant isotype in the resistant AIRmin mice. Additionally, it was shown that IL-4, IL-6, and TNF secreting splenic cells were

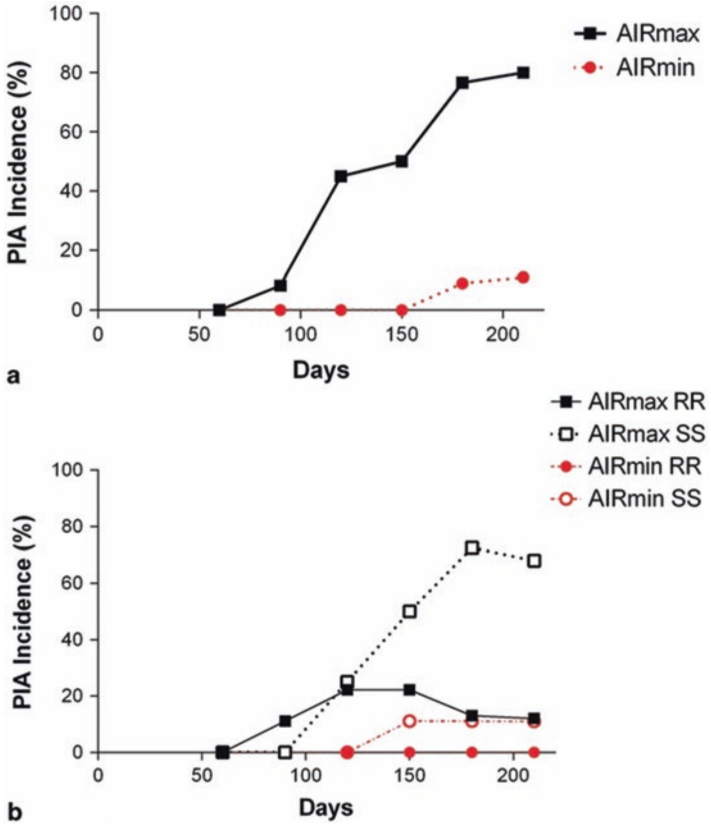


Fig. 12.2 PIA incidence in AIRmax and AIRmin mice and their sublines homozygous for the Slc11a1 gene R and S alleles. Mice received two IP injections of pristane with 60 days interval

significantly more abundant in AIRmax than in AIRmin animals. IFNg-producing cells, on the other hand, increased only in AIRmin mice. Specific pathogen-free susceptible mice do not develop this disease, but when transferred to a conventional environment, they reacquire arthritis susceptibility, indicating the involvement of environmental factors in PIA (Thompson and Elson 1993).

The results in the AIRmax and AIRmin PIA model, when compared to those obtained in inbred mice, evidence the interference of genetic background in the mechanisms underlying arthritis susceptibility and severity. Interaction of arthritis controlling genes with heterogeneous genetic backgrounds and variability in gut microbiota might contribute to the variable signs of arthritis occurring in humans.

The transporter gene *Solute carrier 11 family a1* (*Slc11a1*) has been described in mice as a major modulator of susceptibility to infectious diseases and is expressed in macrophages and neutrophils. *Slc11a1* is pleiotropic, interfering with macrophage activation, oxidative and nitrosamine bursts, TNF, IFNg, and IL-1 production, and the expression of MHC class II molecules. In mice, the mutation corresponding to the *Slc11a1 S* allele associated with susceptibility determines a gly169asp substitution resulting in a nonfunctional protein that promotes an accumulation of ions inside the phagosome of macrophages that favors pathogen replication (Vidal et al. 1992). In the experiment for the production of AIRmax and AIRmin mouse lines, the frequency of the *Slc11a1 S* allele was 25% in the founder population (F0), but shifted to 60% in AIRmin and to 9% in AIRmax after 30 generations of selective breeding. The results suggest that these changes in allele frequencies were the result of the selection process for acute inflammatory response (Araujo et al. 1998).

The effect of the *Slc11a1 R* and *S* alleles during PIA development was studied in AIRmax and AIRmin mice that were rendered homozygous for the *Slc11a1* alleles by genotype-assisted breeding (Fig. 12.2b). AIRmax mice homozygous for the *S* allele (AIRmax^{SS}) were significantly more susceptible (80% incidence) to RA than AIRmax^{RR} mice (30% incidence) evidencing the influence of this polymorphism in RA (Peters et al. 2007). The involvement of this gene in this study as well as in other murine arthritis models constituted the basis for the study of *Slc11a1* involvement in human RA. In fact, several authors reported linkage of *SLC11A1* alleles to human RA probably associated with a polymorphic repeat in the RUNX1-containing promoter region of the gene (Ates et al. 2009).

12.6 Mapping of QTL Controlling PIA in AIRmax and AIRmin Mice

A genome-wide linkage study was carried out in a large F2 population of intercrossed AIRmax and AIRmin F2(AIRmax x AIRmin) mice through linkage analysis of PIA severity phenotype with a panel of SNPs. Two new PIA QTLs (*Prtia* 2 and *Prtia*3) were mapped on chromosomes 5 and 8, respectively, and three suggestive QTLs were detected on chromosomes 7, 17, and 19 (De Franco et al. 2014). In this same F2(AIRmax x AIRmin) population, loci that regulate the intensity of the acute inflammatory response were mapped on chromosomes 5, 7, 8, and 17, which overlap the QTLs that controls PIA severity, suggesting common regulations (Vorraro et al. 2010; Galvan et al. 2011). Co-located chromosome 5 QTLs controlling arthritis severity and humoral responses during *B. burgdorferi* infection were identified in the F2 intercross of C3H/HeNCr and C57BL/6NCr mice (Weis et al. 1999), suggesting the involvement of the chemokine Cxcl9 gene, which maps to the QTL peak in this model (Ma et al. 2009).

In order to candidate genes within the QTL detected in the AIRmax and AIRmin model, transcriptome studies were performed using tissues or cells from normal or arthritic individuals. In this model, the total number of up- and downregulated genes in each line was distinct, as can be seen in Fig. 12.3. More genes were modulated in AIRmax than in AIRmin mice, although a gene ontology analysis revealed an overrepresentation of genes related to inflammatory reaction and chemotaxis biological themes in both lines (Fig. 12.4). Global gene expression analysis indicated 419 differentially

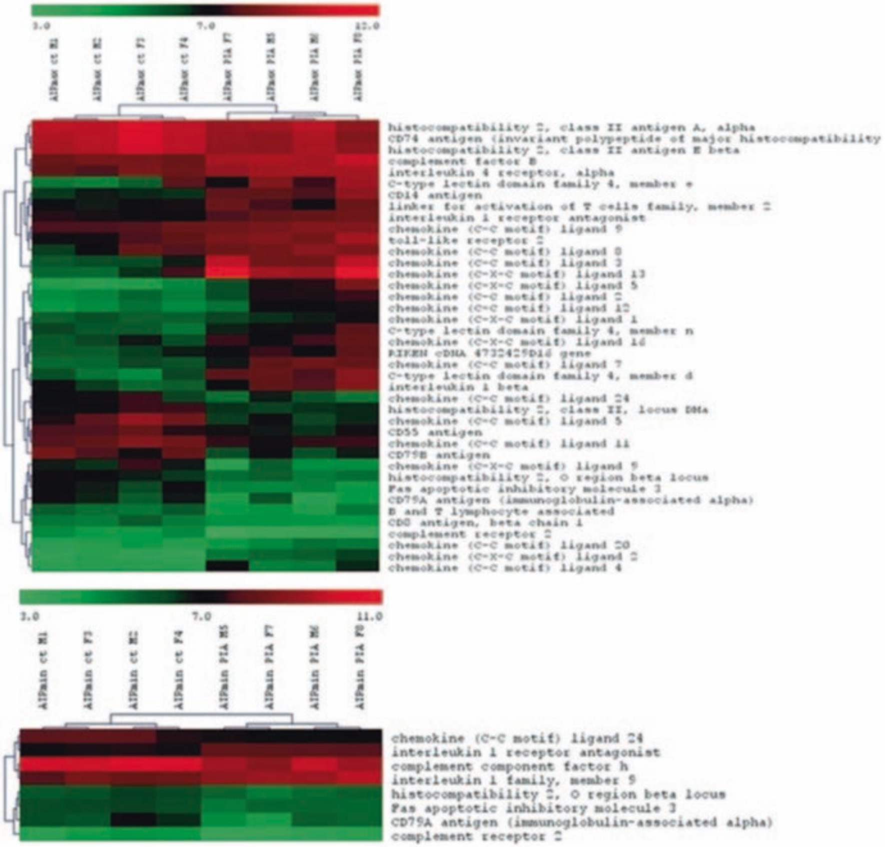


Fig. 12.3 Up- and downmodulated inflammatory and chemokine genes in AIRmax and AIRmin mice. Total RNA was extracted from arthritic paws at 160 days after pristane injection

expressed genes between AIRmax and AIRmin mice. Figs. 12.5 and 12.6 show genes differentially expressed on chromosomes 5 and 8 respectively. Several genes related to inflammation, cell adhesion, and chemotaxis could be observed on chromosome 5, while tissue antigens, cell differentiation, hemeoxigenase, and scavenger receptor genes were observed on chromosome 8 (De Franco et al. 2014).

Ibrahim and collaborators investigated the gene expression profiles of inflamed paws in DBA/1 inbred mice using a similar approach for collagen-induced arthritis (Ibrahim et al. 2002). In their work, inflammation resulted in increased gene expression of matrix metalloproteinases, and immune-related extracellular matrix and cell-adhesion molecules, as well as molecules involved in cell division and transcription, in a manner very similar to the AIRmax/AIRmin model. However, the total number of differentially expressed genes involved in the inbred mice model (223) was lower than in our model (419), suggesting that the heterogeneous background of AIRmax and AIRmin mice permitted a larger genome involvement in this phenotype. Among the differentially expressed genes, inflammatory and chemokine

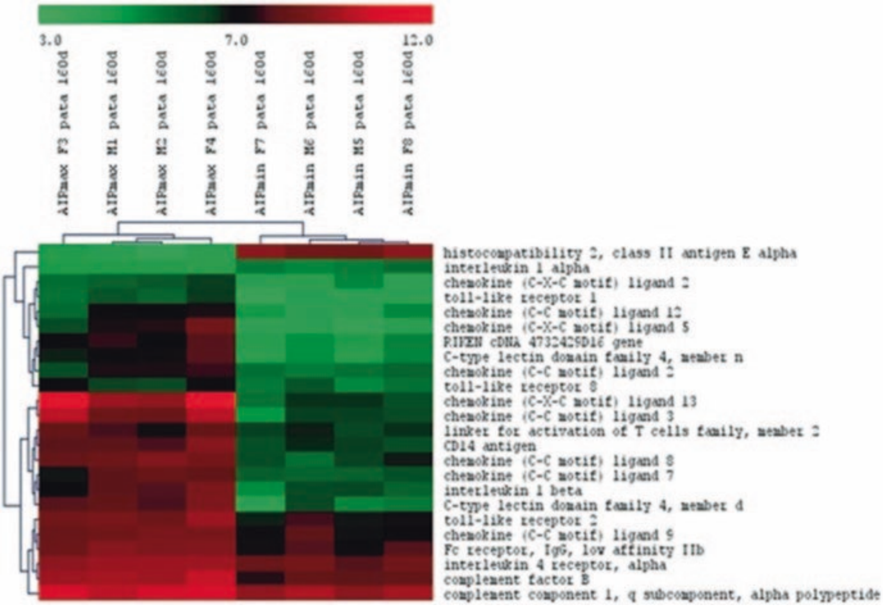


Fig. 12.4 Differentially expressed inflammatory and chemokine genes between AIRmax and AIRmin mice

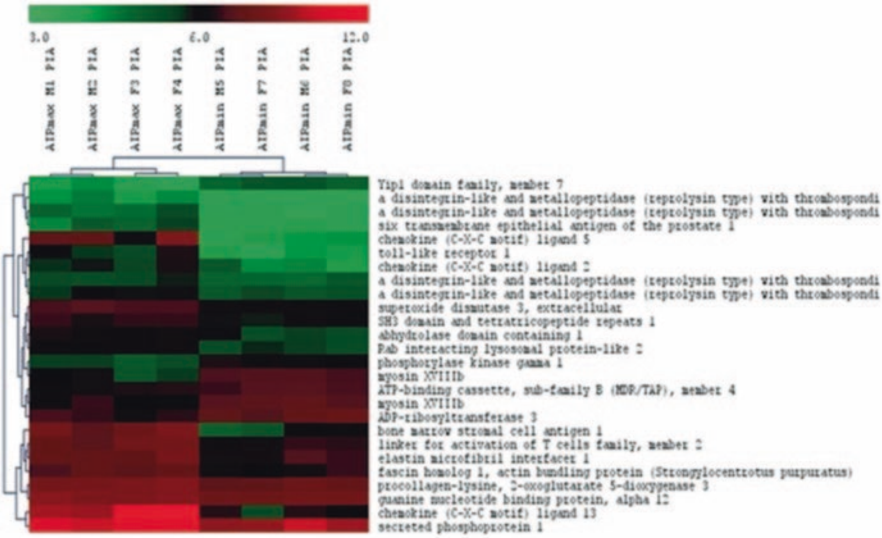


Fig. 12.5 Differentially expressed genes between AIRmax and AIRmin mice mapping at chromosome 5

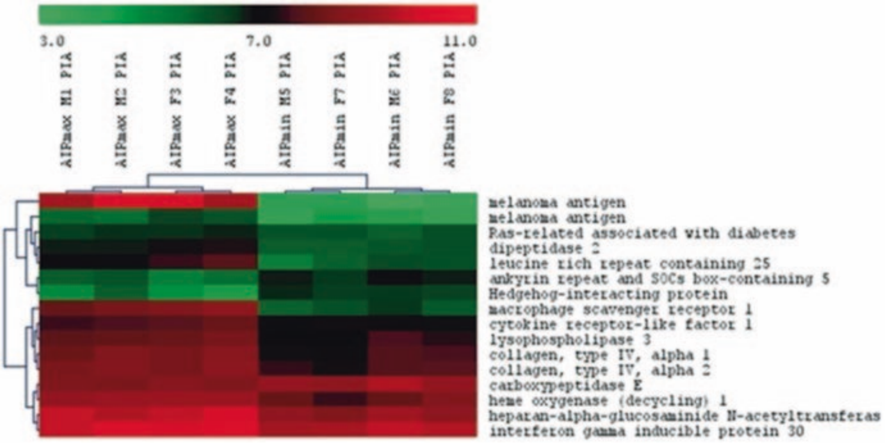


Fig. 12.6 Differentially expressed genes between AIRmax and AIRmin mice mapping at chromosome 8

genes on chromosome 5 and *macrophage scavenger receptor 1* (*Msr1*) and *heme-oxigenase 1* (*Hmox1*) genes on chromosome 8 appear to be the major candidates.

Chemokines are involved in leukocyte recruitment to inflammatory sites, such as synovial tissue in rheumatoid arthritis (RA). However, they may also be homeostatic as these functions often overlap (Ibrahim et al. 2001). Chemokines have essential roles in the recruitment and activation of leucocyte subsets within tissue microenvironments, and stromal cells actively contribute to these networks. Macrophages play a central role in the pathogenesis of rheumatoid arthritis (RA), which is marked by an imbalance of inflammatory and anti-inflammatory macrophages in RA synovium. Although the polarization and heterogeneity of macrophages in RA have not been fully elucidated, the identities of macrophages in RA can potentially be defined by their products, including costimulatory molecules, scavenger receptors, cytokines/chemokines and their receptors, and transcription factors (Li et al. 2012). It has been demonstrated that inappropriate constitutive chemokine expression contributes to the persistence of inflammation by actively blocking its resolution (Filer et al. 2008). This was also observed in urethane-induced lung carcinogenesis, where transcriptome analysis revealed that the genes involved in transendothelial migration and chemokine-cell adhesion were differently expressed in normal lungs of susceptible AIRmin and resistant AIRmax mice (De Franco et al. 2010), suggesting important roles for these phenotypes in chronic diseases.

12.7 MicroRNA and Arthritis

Several studies have demonstrated the involvement of small RNAs, known as miRNAs in the development of RA. MiRNAs are a class of small, noncoding, RNA molecules with approximately 21 nucleotides in length that can regulate gene expression by reducing the ability of specific mRNAs to direct the synthesis of their encoded proteins (Krol et al. 2010). They likely participate in most developmental and physiologic processes, with involvement in, but not limited to, cell proliferation and differentiation, regulation of lipid metabolism, and modulation of insulin secretion. The importance of miRNA-mediated regulation of gene expression for the prevention of autoimmunity and maintenance of normal immune system functions has been described (Wittmann and Jäck 2011). Studies in humans have detected altered miRNA expression in RA patients when compared to controls or osteoarthritis patients (Pauley et al. 2008; Kobayashi et al. 2008; Stanczyk et al. 2008). MiRNAs can be detected in body fluids without invasive procedures, and thus may be used as prognostic or diagnostic biomarkers for specific conditions, such as rheumatic diseases (Ceribelli et al. 2012).

In the PIA model, pristane injection modulated several genes in the peritoneal cells of AIRmax and AIRmin lines in both time points analyzed. This modulation was more widespread in AIRmax mice, with about twice the number of modulated genes than the AIRmin line (2025 vs 1043). This difference reflects mainly the number of downregulated genes, which was five-fold higher in AIRmax animals (704 vs 131). In previous microarray analyses using the paws of these animals, the AIRmax line also showed five-fold more downregulated genes than AIRmin and two-fold more upregulated genes (De Franco et al. 2014). The same gene expression profile was also observed in the subcutaneous tissue of these lines after Biogel injection (Fernandes et al. 2016). Although different tissues and stimuli have been analyzed, these results indicate that the selective pressure during phenotypic selection acted in general inflammatory regulation mechanisms.

MiRNA expressions after pristane injection were also distinct in AIRmax and AIRmin mice. At 120 days, 184 miRNAs were upregulated and 12 downregulated exclusively in AIRmax animals. That regulation was similar (189 up- and 12 downregulated) at 170 days. In contrast, the AIRmin line upregulated 15 and 10 miRNAs at 120 and 170 days respectively; no downregulated miRNA was detected. The higher number of downregulated genes observed in AIRmax mice may be a consequence of the upregulation of miRNAs in their peritoneal cells.

Most of the up- or downregulated miRNAs have not been ascribed roles in experimental or human arthritis development. Instead, many of those miRNAs are described in terms of their roles in suppressing or inducing several types of malignant tumors, although many have been shown to be involved in the regulation of important biological processes in the development of autoimmune diseases such as inflammation. We therefore sought to identify important pathways in which those miRNAs participated and which could explain their modulation in our model – eventually leading to the identification of new arthritis-related miRNAs in experimentally induced arthritis.

MiR-132-3p was the most upregulated miRNA in the susceptible mouse line in microarrays and qRT-PCR (106- and seven-fold higher at 120 days, and 67- and 4.5-fold higher at 170 days respectively). Expression of that miRNA has been found to increase in the peripheral blood mononuclear cells (PBMCs) of rheumatoid arthritis patients (Kobayashi et al. 2008). In that study, one of the patients with the active disease showed unaltered levels of that and other miRNAs related to the disease after two months of treatment with methotrexate. Those results indicate that the high expressions of miRNAs in that patient were related to unresponsiveness to the treatment. MiR-132-3p may therefore play a key role in systemic conditions related to joint inflammation, which would explain its high expression in the peritoneum of susceptible AIRmax animals. MiR-132 is specifically induced in Th17 cells and acts as a proinflammatory mediator increasing osteoclastogenesis through the downregulation of COX2. In in vivo, articular knockdown of MiR-132 in murine arthritis models reduces the number of osteoclasts in the joints (Donate et al. 2021).

MiR-132-3p and miR-212-3p are members of the same family (located on chromosome 11 in mice) that forms the miR-212/132 cluster, and they have similar seed sequences. That cluster, induced by the activation of AhR in inflammatory bowel disease, was able to promote an inflammatory response by inducing the Th17 response and suppressing IL-10 production (Chinen et al. 2015). The *Il10* gene was downregulated in peritoneal cells in AIRmax mice, indicating that there may be an indirect regulation of the expression of that cytokine by those miRNAs (Fernandes et al. 2018). IL-10 is an important anti-inflammatory cytokine that inhibits proinflammatory mediator production and lymphocyte proliferation, thus playing a protective role in autoimmune diseases. IL-10 has been shown to contribute to the prevention of arthritic inflammation in macrophages during collagen-induced arthritis development (Chen et al. 2017). That gene can be regulated by different miRNAs, including miR-27b-3p, which is highly upregulated in that line (Fig. 12.7).

Cd69 and S1pr1 (specifically targeted by 106a-5p, 25-3p, and 20b-5p miRNAs) were downregulated in AIRmax mice (Fig. 12.7). CD69 is a leukocyte receptor induced in lymphocytes and macrophages after activation. Sancho et al. 2003 demonstrated that CD69−/− and CD69 +/− mice had an exacerbated form of collagen-induced arthritis (CIA) when compared to controls and that CD69 was capable of inducing TGF-β2 synthesis. TGF-β2 is an anti-inflammatory cytokine, and null mutations in that gene can lead to severe inflammatory disorders; that gene regulates the production of inflammatory cytokines and has protective effects in the CIA model (Sancho et al. 2003; Brandes et al. 1991). *Tgfb2* was the most downregulated gene in the AIRmax line (40-fold), while CD69 was approximately six-fold downregulated. The *S1pr1* gene, on the other hand, affects the differentiation of osteoblasts (Sato et al. 2012). The inhibition of osteoblast differentiation contributes to bone loss in RA as well as to a decreased healing ability of those lesions (Baum and Gravallese 2016).

The expressions of miR-181b-5p and *Il6* were shown to be inversely correlated following stimulation with LPS, and *Il6* is a direct target of miR-181b-5p (Zhang et al. 2015), demonstrating the critical role of the posttranscriptional control of IL-6 by miR-181b-5p in endotoxin tolerance. The expressions of miR-181b-5p and *Il6* were also inversely correlated in susceptible AIRmax mice. Although *Il6* did not appear as a target for miR-181b-5p in our interaction network (which considered at least 3 different algorithms), the data from the TargetScan database (which is widely used in the literature to predict miRNA-RNA interactions) indicated that gene as a possible target of miR-181b-5p. An important role of IL-6 has been reported in the

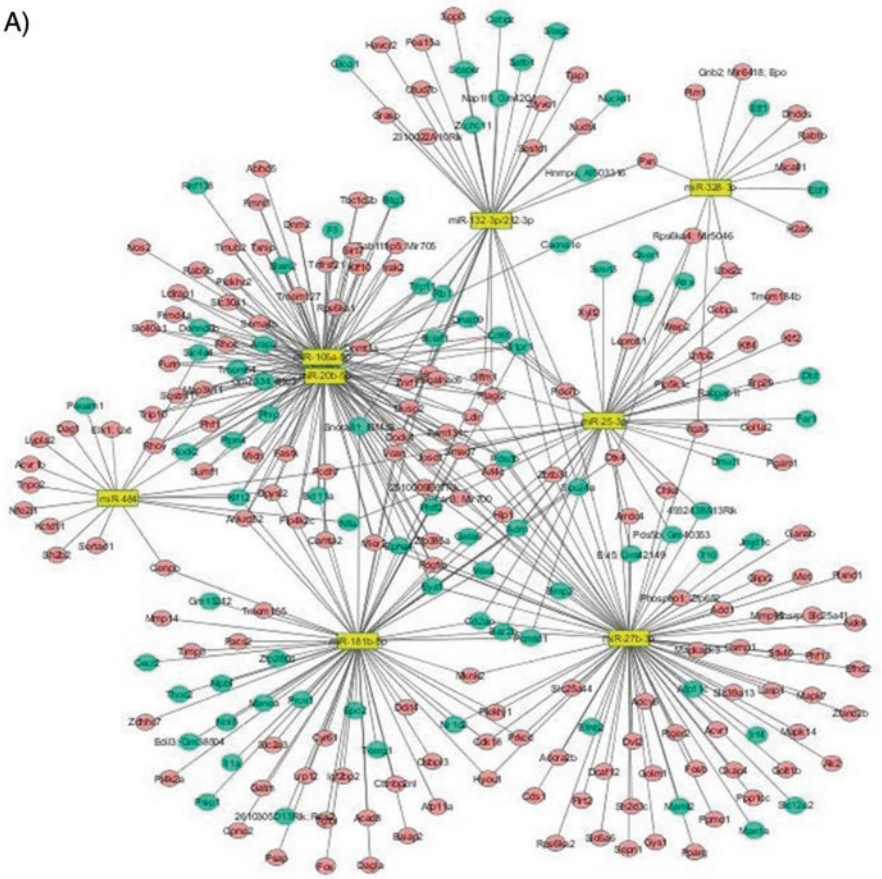


Fig. 12.7 mRNA-miRNA interaction network. (**a**) miRNAs upregulated in AIRmax mice and their interaction with predicted target genes; (**b**) miRNAs upregulated AIRmin mice and their interaction with predicted target genes. Red = upregulated genes; green = downregulated genes. The interaction network was built with Cytoscape 3.4.0.

in vitro inhibition of osteoclast progenitors mediated by the disruption of RANK signaling (Yoshitake et al. 2008). Osteoclasts are required for articular bone resorption and are responsible for bone erosion in RA (Baum and Gravallese 2016; Lin et al. 2015). The unbalanced expression of the genes that promote osteoclatogenesis and inhibit osteoblast differentiation may represent a mechanism for the stimulation of bone erosion and increased disease severity in AIRmax animals. Histological analyses of the AIRmax paws did, in fact, show bone loss in addition to the destruction of cartilage (Correa et al. 2017).

Soto et al. 2008 compared the gene expression profiles of the rat collagen-induced arthritis model (CIA) with human RA (using paw and knee synovial tissue respectively). Comparing the DEGs in our model with the model used by Soto, we

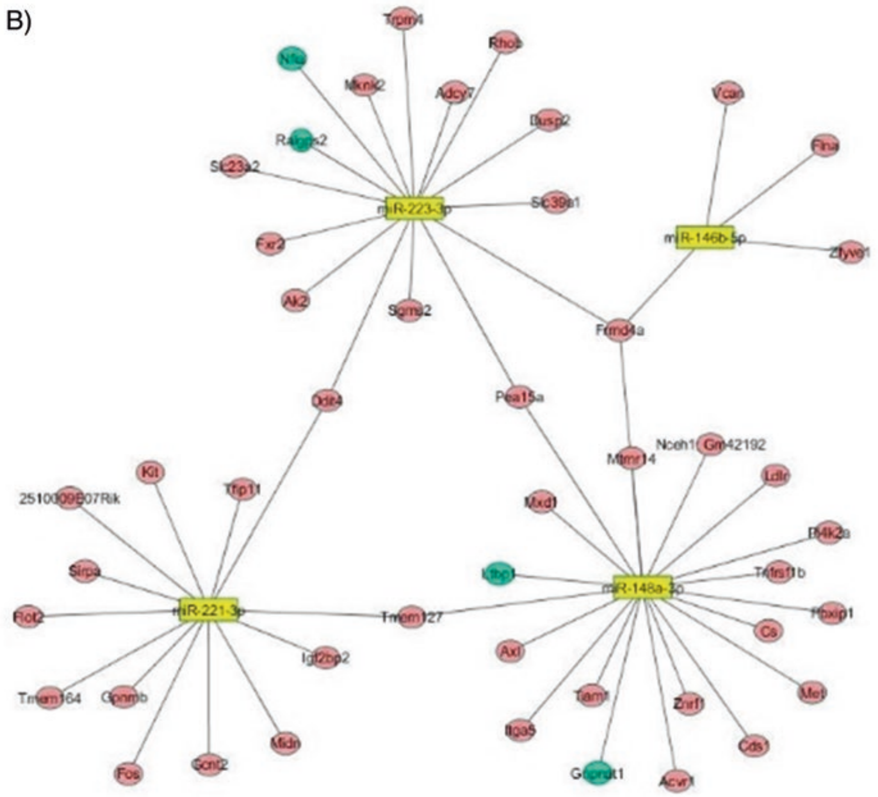


Fig. 12.7 (continued)

observed that two genes upregulated in AIRmax mice (Mmp13 and Gpsm3) were also upregulated in CIA rats.

The *MMP13* and *GPSM3* genes play significant roles in rheumatoid arthritis in humans, and *GPSM3* has been associated with the risk of developing autoimmune diseases. Polymorphisms associated with decreased transcription have been inversely correlated with the risk of developing arthritis. The reduced expression of *GPSM3* was observed to prevent neutrophil migration mediated by LTB4 (leukotriene B4) and CXCL8 to arthritic joints (Gall et al. 2016). Additionally, mice deficient for Gpsm3 were protected from arthritis induced by anticollagen antibodies, with reduced CCL2- and CX3CL1-mediated migration of myeloid cells (Giguère et al. 2013). *Gpsm3* is located on chromosome 17 in mice, where a suggestive QTL for experimental arthritis was detected in our model (De Franco et al. 2014). The miRanda database identified *Gpsm3* as a predicted target of miRNA-151-5p, which is downregulated in AIRmax mice. Since the interaction was only predicted by the database, it was not considered in our results, although the high expression of *Gpsm3* as a consequence of the downregulation of miRNA-151-3p should not be completely ruled out.

MMP-13 (or collagenase-3) hydrolyzes type 2 collagen and may favor the destruction of cartilage in arthritic joints. In rheumatoid arthritis, IL-1β and TNF-α produced by macrophages in the connective tissue stimulate the production of that MMP by articular chondrocytes (Vincenti and Brinckerhoff 2002). Additionally, a key role has been attributed to some genetic loci encoding metalloproteinases in bone destruction. The expression of MMP-13 increased ten-fold in AIRmax mice but remained unaltered in pristane-treated AIRmin animals. Vonk and coworkers (Vonk et al. 2014) looked for different miRNAs expressed in healthy and osteoarthritis (OA) patients and found that miRNA-148a levels in healthy subjects were approximately ten-fold higher than those seen in patients with the disease. Transfection of miR-148a-3p into cells of OA patients resulted in decreased MMP-13 expression (which had increased in those patients), suggesting that this miRNA may play a protective role in OA, with a consequent reduction in cartilage destruction.

In a second analysis, Soto et al. 2008 identified 30 differentially expressed genes when comparing RA patients and healthy controls. Of those 30 genes, *Pde3b*, *Tgfb2*, and *Fam120c* were downregulated in both RA patients and AIRmax mice; *Tgfb2* showed a significant protective effect in arthritis models as discussed above.

Many miRNAs are over- or underexpressed in autoimmune diseases such as SLE (Liang and Shen 2012; Amarilyo and La Cava 2012) and rheumatoid arthritis (RA) (Ceribelli et al. 2011), and investigators have reported that miR-146a is altered in those diseases (Ceribelli et al. 2012). Interestingly, expression of miR-146a was higher in AIRmax than in AIRmin control mice 120 days after pristane injection. Increased expression of miRNA-146a has been well documented in the PBMCs of arthritic patients. That microRNA has two known targets: *Traf6* (TNF receptor-associated factor 6) and *Irak1* (interleukin-1 receptor-associated kinase 1), both of which stimulate TNF-α production (Shrivastava and Pandey 2013). The expression of those molecules were unaltered in those patients, suggesting that increased miRNA-146a levels were unable to regulate TRAF6/IRAK. Therefore, it is not exactly known how the high expression of that miRNA is related to the increased levels of TNF-α in RA (Ceribelli et al. 2011).

12.8 Concluding Remarks

Recent advances in the field of genetics have dramatically changed our understanding of autoimmune disease. Candidate gene and, more recently, genome-wide association (GWA) and linkage studies have led to an explosion in the number of loci and pathways known to contribute to autoimmune phenotypes, confirming a major role for the MHC region and, more recently, identifying risk loci involving both the innate and adaptive immune responses. However, most regions found through GWA scans have yet to isolate the association to the causal allele(s) responsible for conferring disease risk. A role for rare variants (allele frequencies of <1%) has begun to emerge. The study of the abundant long intergenic noncoding RNAs and of small interfering RNA (microRNAs) has also become a powerful tool to understand the mechanisms that modulate the gene expression profiles in RA and other autoimmune diseases (Jarvis and Frank 2010; Donate et al. 2013). Future research will also use next generation sequencing (NGS) technology to comprehensively evaluate the human genome for risk variants. Whole transcriptome sequencing (e.g., RNA-Seq), which combines gene expression, sequence, and splice variant analysis, will provide much more detailed gene expression data. Despite its high incidence and severe phenotype, RA still has no cure in spite of many efforts to produce effective therapy treatments. Further studies should therefore be carried out to better understand the functions and mechanisms of miRNAs in the immune system and in arthritis development. The AIRmax and AIRmin lines constitute interesting tools for mapping inflammatory disease modifying genes and miRNAs, in addition to being a valid animal model for the human disease in respect to similar gene pathways and miRNAs. Our studies have been demonstrated that those lines have distinct gene and miRNA expression profiles, which may be partly responsible for their different phenotypes. Regardless of the current or future technology, the versatility of murine models will continue to be required to advance our understanding of human diseases.

References

Adarichev VA, Valdez JC, Bardos T et al (2003) Combined autoimmune models of arthritis reveal shared and independent qualitative (binary) and quantitative trait loci. J Immunol 170(5):2283–2292

Amarilyo G, La Cava A (2012) miRNA in systemic lupus erythematosus. Clin Immunol 144(1):26–31

Araujo LM, Ribeiro OG, Siqueira M et al (1998) Innate resistance to infection by intracellular bacterial pathogens differs in mice selected for maximal or minimal acute inflammatory response. Eur J Immunol 28(9):2913–2920

Asquith DL, Miller AM, Mcinnes IB et al (2009) Animal models of rheumatoid arthritis. Eur J Immunol 39(8):2040–2044

Ates O, Dalyan L, Musellim B et al (2009) NRAMP1 (SLC11A1) gene polymorphisms that correlate with autoimmune versus infectious disease susceptibility in tuberculosis and rheumatoid arthritis. Int J Immunogenet 36(1):15–19

Baechler EC, Batliwalla FM, Reed AM et al (2006) Gene expression profiling in human autoimmunity. Immunol Rev 210:120–137

Besenyei T, Kadar A, Tryniszewska B et al (2012) Non-MHC risk alleles in rheumatoid arthritis and in the syntenic chromosome regions of corresponding animal models. Clin Dev Immunol. https://doi.org/10.1155/2012/284751

Baum R, Gravallese EM (2016) Bone as a target organ in rheumatic disease: impact on osteoclasts and osteoblasts. Clin Rev Allergy Immunol 51(1):1–15

Biozzi G, Ribeiro OG, Saran A et al (1998) Effect of genetic modification of acute inflammatory responsiveness on tumorigenesis in the mouse. Carcinogenesis 19(2):337–346

Brandes ME, Allen JB, Ogawa Y et al (1991) Transforming Growth Factor Beta 1 suppresses acute and chronic arthritis in experimental animals. J Clin Invest 87(3):1108–1113

Carr HL, Turner JD, Major T, Scheel-Toellner D, Filer A (2020) New developments in transcriptomic analysis of synovial tissue. Front Med 7:21. https://doi.org/10.3389/fmed.2020.00021

Ceribelli A, Ma N, Satoh M et al (2011) MicroRNAs in rheumatoid arthritis. FEBS Lett 585(23):3667–3674

Ceribelli A, Satoh M, Chan EK (2012) microRNAs and autoimmunity. Curr Opin Immunol 24(6):686–691

Chen S, Chen B, Wen Z et al (2017) Il-33/st2-mediated inflammation in macrophages is directly abrogated by Il-10 during rheumatoid arthritis. Oncotarget 8(20):32407–324018

Chinen I, Nakahama T, Kimura A et al (2015) the aryl hydrocarbon receptor/microrna-212/132 axis in t cells regulates il-10 production to maintain intestinal homeostasis. Int Immunol 27(8):405–415

Cho YG, Cho ML, Min SY et al (2007) Type II collagen autoimmunity in a mouse model of human rheumatoid arthritis. Autoimmun Rev 7(1):65–70

Correa MA, Canhamero T, Borrego A et al (2017) Slc11a1 (Nramp-1) gene modulates immune-inflammation genes in macrophages during pristane-induced arthritis in mice. Inflamm Res 66:969–980

De Franco M, Colombo F, Galvan A et al (2010) Transcriptome of normal lung distinguishes mouse lines with different susceptibility to inflammation and to lung tumorigenesis. Cancer Lett 294(2):187–194

De Franco M, Peters LC, Correa MA et al (2014) Pristane-induced arthritis loci interact with the *Slc11a1* gene to determine susceptibility in mice selected for high inflammation. PLoS One 9(2):e88302

Donate PB, Fornari TA, Macedo C et al (2013) T cell post-transcriptional miRNA-mRNA interaction networks identify targets associated with susceptibility/resistance to collagen-induced arthritis. PLoS One 8(1):e54803

Donate PB, Alves de Lima K, Peres RS, et al (2021) Cigarette smoke induces *miR-132* in Th17 cells that enhance osteoclastogenesis in inflammatory arthritis. Proc Natl Acad Sci USA 5;118(1):e2017120118

Fernandes JG, Canhamero T, Borrego A, et al (2016) Distinct gene expression profiles provoked by polyacrylamide beads (Biogel) during chronic and acute inflammation in mice selected for maximal and minimal inflammatory responses. Inflamm Res 2016;65(4):313–323

Fernandes JG, Borrego A, Jensen JR et al (2018) miRNA Expression and Interaction with Genes Involved in Susceptibility to Pristane-Induced Arthritis. J Immunol Res Dec 16:1928405. https://doi.org/10.1155/2018/1928405

Filer A, Raza K, Salmon M et al (2008) The role of chemokines in leucocyte-stromal interactions in rheumatoid arthritis. Front Biosci 13:2674–2685

Fujikado N, Saijo S, Iwakura Y (2006) Identification of arthritis-related gene clusters by microarray analysis of two independent mouse models for rheumatoid arthritis. Arthritis Res Ther 8(4):100–125

Gall BJ, Schroer AB, Gross JD et al (2016) Reduction of *Gpsm3* expression skin to the arthritis-protective SNP rs204989 differentially affects migration in a neutrophil model. Genes Immun 17(6):321–327

Galvan A, Vorraro F, Cabrera W et al (2011) Association study by genetic clustering detects multiple inflammatory response loci in non-inbred mice. Genes Immun 12(5):390–394

Giguère PM, Billard MJ, Laroche G et al (2013) G-protein signaling modulator-3, a gene linked to autoimmune diseases, regulates monocyte function and its deficiency protects from inflammatory arthritis. Mol Immunol 54(2):193–198

Glant TT, Finnegan A, Mikecz K (2003) Proteoglycan-induced arthritis: immune regulation, cellular mechanisms, and genetics. Crit Rev Immunol 23(3):199–250

Glant TT, Adarichev VA, Nesterovitch AB et al (2004) Disease-associated qualitative and quantitative trait loci in proteoglycan-induced arthritis and collagen-induced arthritis. Am J Med Sci 327(4):188–195

Hirose J, Tanaka S (2011) Animal models for bone and joint disease. CIA, CAIA model. Clin Calcium 21(2):253–259

Ibanez OM, Stiffel C, Ribeiro OG et al (1992) Genetics of nonspecific immunity: I. Bidirectional selective breeding of lines of mice endowed with maximal or minimal inflammatory responsiveness. Eur J Immunol 22(10):2555–2563

Ibrahim SM, Yu X (2006) Dissecting the genetic basis of rheumatoid arthritis in mouse models. Curr Pharm Des 12(29):3753–3759

Ibrahim SM, Mix E, Bottcher T et al (2001) Gene expression profiling of the nervous system in murine experimental autoimmune encephalomyelitis. Brain 124:1927–1938

Ibrahim SM, Koczan D, Thiesen HJ (2002) Gene-expression profile of collagen-induced arthritis. J Autoimmun 18(2):159–167

Jarvis JN, Frank MB (2010) Functional genomics and rheumatoid arthritis: where have we been and where should we go? Genome Med 2(7):44–59

Jensen JR, Peters LC, Borrego A et al (2006) Involvement of antibody production quantitative trait loci in the susceptibility to pristane-induced arthritis in the mouse. Genes Immun 7(1):44–50

Kannan K, Ortmann RA, Kimpel D (2005) Animal models of rheumatoid arthritis and their relevance to human disease. Pathophysiology 12(3):167–181

Kobayashi T, Lu J, Cobb BS et al (2008) Dicer-dependent pathways regulate chondrocyte proliferation and differentiation. Proc Natl Acad Sci U S A 105(6):1949–1954

Kobezda T, Ghassemi-Nejad S, Mikecz K et al (2014) Of mice and men: how animal models advance our understanding of T-cell function in RA. Nat Rev Rheumatol 10(3):160–170

Krol J, Loedige I, Filipowicz W (2010) The widespread regulation of microrna biogenesis, function and decay. Nat Rev Genet 11(9):597–610

Kurko J, Besenyei T, Laki J et al (2013) Genetics of rheumatoid arthritis—a comprehensive review. Clin Rev Allergy Immunol 45(2):170–179

Li P, Schwarz EM (2003) The TNF-alpha transgenic mouse model of inflammatory arthritis. Springer Semin Immunopathol 25(1):19–33

Li J, Hsu HC, Mountz JD (2012) Managing macrophages in rheumatoid arthritis by reform or removal. Curr Rheumatol Rep 14(5):445–454

Liang D, Shen N (2012) MicroRNA involvement in lupus: the beginning of a new tale. Curr Opin Rheumatol 24(5):489–498

Lin L, Yee SW, Kim RB et al (2015) Slc transporters as therapeutic targets: emerging opportunities. Nat Rev Drug Discov 14(8):543–560

Lubberts E, Koenders MI, Van Den Berg WB (2005) The role of T-cell interleukin-17 in conducting destructive arthritis: lessons from animal models. Arthritis Res Ther 7(1):29–37

Ma Y, Miller JC, Crandall H et al (2009) Interval-specific congenic lines reveal quantitative trait Loci with penetrant lyme arthritis phenotypes on chromosomes 5, 11, and 12. Infect Immun 77(8):3302–3311

Pauley KM, Satoh M, Chan AL et al (2008) Upregulated mir-146a expression in peripheral blood mononuclear cells from rheumatoid arthritis patients. Arthritis Res Ther 10(4):r101

Patten C, Bush K, Rioja I et al (2004) Characterization of pristane-induced arthritis, a murine model of chronic disease: response to antirheumatic agents, expression of joint cytokines, and immunopathology. Arthritis Rheum 50(10):3334–3345

Peters LC, Jensen JR, Borrego A et al (2007) Slc11a1 (formerly NRAMP1) gene modulates both acute inflammatory reactions and pristane-induced arthritis in mice. Genes Immun 8(1):51–56

Potter M, Wax JS (1981) Genetics of susceptibility to pristane-induced plasmacytomas in BALB/cAn: reduced susceptibility in BALB/cJ with a brief description of pristane-induced arthritis. J Immunol 127(4):1591–1595

Roper RJ, Weis JJ, Mccracken BA et al (2001) Genetic control of susceptibility to experimental Lyme arthritis is polygenic and exhibits consistent linkage to multiple loci on chromosome 5 in four independent mouse crosses. Genes Immun 2(7):388–397

Runstadler JA, Saila H, Savolainen A et al (2005) Association of SLC11A1 (NRAMP1) with persistent oligoarticular and polyarticular rheumatoid factor-negative juvenile idiopathic arthritis in Finnish patients: haplotype analysis in Finnish families. Arthritis Rheum 52(1):247–256

Sancho D, Gómez M, Viedma F et al (2003) CD69 downregulates autoimmune reactivity through active transforming growth factor-beta production in collagen-induced arthritis. J Clin Invest 112(6):872–882

Sato C, Iwasaki T, Kitano S et al (2012) Sphingosine 1-phosphate receptor activation enhances bmp-2-induced osteoblast differentiation. Biochem Biophys Res Commun 423(1):200–205

Silman AJ, Pearson JE (2002) Epidemiology and genetics of rheumatoid arthritis. Arthritis Res 4(Suppl 3):S265–S272

Silva GL, Junta CM, Sakamoto-Hojo ET et al (2009) Genetic susceptibility loci in rheumatoid arthritis establish transcriptional regulatory networks with other genes. Ann N Y Acad Sci 1173:521–537

Shrivastava AK, Pandey A (2013) Inflammation and rheumatoid arthritis. J Physiol Biochem 69(2):335–347

Soto H, Hevezi P, Roth RB et al (2008) Gene array analysis comparison between rat collagen-induced arthritis and human rheumatoid arthritis. Scand J Immunol 68(1):43–57

Stanczyk J, Pedrioli DM, Brentano F et al (2008) Altered expression of microRNA in synovial fibroblasts and synovial tissue in rheumatoid arthritis. Arthritis Rheum 58(4):1001–1009

Stasiuk LM, Ghoraishian M, Elson CJ et al (1997) Pristane-induced arthritis is CD4+ T-cell dependent. Immunology 90(1):81–86

Teixeira VH, Olaso R, Martin-Magniette ML et al (2009) Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. PLoS One 4(8):e6803

Thompson SJ, Elson CJ (1993) Susceptibility to pristane-induced arthritis is altered with changes in bowel flora. Immunol Lett 36(2):227–231

Thompson SJ, Rook GA, Brealey RJ et al (1990) Autoimmune reactions to heat-shock proteins in pristane-induced arthritis. Eur J Immunol 20(11):2479–2484

Thompson SJ, Hitsumoto Y, Zhang YW et al (1992) Agalactosyl IgG in pristane-induced arthritis. Pregnancy affects the incidence and severity of arthritis and the glycosylation status of IgG. Clin Exp Immunol 89(3):434–438

Thompson SJ, Francis JN, Siew LK et al (1998) An immunodominant epitope from mycobacterial 65-kDa heat shock protein protects against pristane-induced arthritis. J Immunol 160(9):4628–4634

Van Den Berg WB (2009) Lessons from animal models of arthritis over the past decade. Arthritis Res Ther 11(5):250–259

Vidal SM, Epstein DJ, Malo D et al (1992) Identification and mapping of six microdissected genomic DNA probes to the proximal region of mouse chromosome 1. Genomics 14(1):32–37

Vincenti MP, Brinckerhoff CE (2002) Transcriptional regulation of collagenase (*Mmp-1, Mmp-13*) genes in arthritis: integration of complex signaling pathways for the recruitment of gene-specific transcription factors. Arthritis Res 4(3):157–164

Vigar ND, Cabrera WH, Araujo LM et al (2000) Pristane-induced arthritis in mice selected for maximal or minimal acute inflammatory reaction. Eur J Immunol 30(2):431–437

Vonk LA, Kragten AH, Dhert WJ et al (2014) Overexpression of Hsa-Mir-148a promotes cartilage production and inhibits cartilage degradation by osteoarthritic chondrocytes. Osteoarthr Cartil 22(1):145–153

Vorraro F, Galvan A, Cabrera WH et al (2010) Genetic control of IL-1 beta production and inflammatory response by the mouse Irm1 locus. J Immunol 185(3):1616–1621

Weis JJ, Mccracken BA, Ma Y et al (1999) Identification of quantitative trait loci governing arthritis severity and humoral responses in the murine model of Lyme disease. J Immunol 162(2):948–956

Wittmann J, Jäck HM (2011) MicroRNAs in rheumatoid arthritis: midget rnas with a giant impact. Ann Rheum Dis 70(suppl 1):i92–i96

Yoshitake F, Itoh S, Narita H et al (2008) Interleukin-6 directly inhibits osteoclast differentiation by suppressing receptor activator of nf-kappab signaling pathways. J Biol Chem 283(17):11535–11540

You S, Yoo SA, Choi S et al (2014) Identification of key regulators for the migration and invasion of rheumatoid synoviocytes through a systems approach. Proc Natl Acad Sci U S A 111(1):550–555

Yu X, Bauer K, Koczan D et al (2007) Combining global genome and transcriptome approaches to identify the candidate genes of small-effect quantitative trait loci in collagen-induced arthritis. Arthritis Res Ther 9(1):3–17

Zhang W, Shen X, Xie L et al (2015) MicroRNA-181b regulates endotoxin tolerance by targeting il-6 in macrophage raw264.7 cells. J Inflamm 12:18. https://doi.org/10.1186/s12950-015-0061-8

Chapter 13
Trannscriptomics and Immune Response in Human Cancer

L. P. Chaves, C. M. Melo, W. Lautert-Dutra, A. L. Caliari, and J. A. Squire

13.1 Introduction

Cancer transcriptomics uses high-throughput methods to determine the abundance and relative expression levels of every active gene in a tumor (Cieślik and Chinnaiyan 2018). The actively transcribed RNA in cancer is highly dynamic, reflecting the tissue of origin of the tumor, disrupted regulatory mechanisms of cancer genes, and tumor–host interactions. The cancer transcriptome profile can be regarded as a gene expression signature of the underlying cell state of the diverse population of cells at the time of tumor sampling. Thus, transcriptomic profiling of patient tumors can provide experimental approaches to defining molecular pathways that have been activated and are responsible for driving the cancer process.

Changes in gene transcriptional activity are known to mediate various tumor phenotypes, such as inflammation, vascularization, apoptosis, proliferation, immune evasion, and genomic instability that are considered the hallmarks of cancer progression (Hanahan and Weinberg 2011). In addition, to identifying driver pathways of oncogenesis, transcriptomic analysis of tumor RNA also contains a wealth of gene expression information related to microenvironmental interactions that impact immunotherapy responses (Galon and Bruni 2019).

Gene expression profiling of cancer started with the application of microarray analysis to study tumor-specific patterns of gene expression (reviewed in Macgregor and Squire 2002). There are many different commercial microarray platforms, but in typical gene microarray design, several thousand known gene probe sequences are orderly attached to the solid surface of the array. Fluorescent molecules are labeled on cDNA copies of the total RNA from the tumor sample, which are then

L. P. Chaves · C. M. Melo · W. Lautert-Dutra · A. L. Caliari · J. A. Squire (✉)
Department of Genetics, Ribeirão Preto Medical School, University of São Paulo,
Ribeirão Preto, SP, Brazil
e-mail: squirej@fmrp.usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_13

hybridized with the fixed gene probes on the array. According to the amount of hybridization between the tagged tumor cDNA and the fixed gene probes bound to the array, the fluorescence intensity varies and is converted into data that indicates expression information for each gene probe. This early profiling platform provided valuable insights concerning the complexity of gene expression in cancer biology and paved the way for molecular classifications of clinical subtypes (discussed in the next section) (Colombo et al. 2011).

The ease of use and relatively inexpensive cost of microarrays encouraged their popularity for the study of cancer transcriptomics. One major limitation of all the microarray platforms is that the coverage of the target gene probes depends upon existing knowledge about the genes or transcripts being studied. Another disadvantage of all array methods is the high background levels due to cross-hybridization between related genes. In addition to these issues, the narrow dynamic range of detection due to both background and saturation signals means that the platform's sensitivity has been a significant limitation. For these reasons, current microarray platforms are unlikely to identify gene expression differences arising from tumor samples containing mixed populations of cells.

In recent years, RNA sequencing (RNA-seq) has emerged as a more sensitive and versatile alternative for gene expression profiling (Wang et al. 2009). Since RNA-seq is not limited to profiling predefined transcripts/genes, it is able to provide a complete overview of the expression of the whole transcriptome. This comprehensive coverage means that RNA-seq can identify more differentially modulated transcripts of relevance to cancer, splice variants, and noncoding transcripts [e.g., microRNA (miRNA), long noncoding RNA (lncRNA), pseudogenes]. These additional data may be informative for improved molecular classifications of the cancer transcriptome in the context of mutational data from DNA sequencing, mechanistic investigations, and biomarker discovery (Malone et al. 2020).

Cancer transcriptome profiling has been greatly aided by bioinformatics methods that enable researchers to link the somatic mutations in a tumor with clinical phenotypes such as drug response or overall survival. The functional phenotypes that can be interrogated through transcriptome profiling are very broad and include quantitative estimates of expression levels and the detection of transcript isoforms, fusion RNAs, RNA-editing sites, and noncoding RNA (Uhlen et al. 2017). Beyond tumor cell-intrinsic features, transcriptome profiling can provide insights into the tumor microenvironment, for example, by characterizing transcripts from different types of infiltrating T cells during an immune response.

This chapter will address current transcriptome research and translational approaches using gene expression data to improve understanding of immune responses against cancer. We will focus on the common bioinformatics approaches and gene expression databases that new researchers and translational oncologists will find helpful for initiating studies on tumor transcriptomes. We discuss *in silico* and single-cell RNA-seq methods currently used to determine the number and type of different immune cells in a tumor, and we will briefly consider emerging new directions in this field.

13.2 Gene Expression Approach for Clinical Investigations

Whole exome sequencing (WES) uses coding regions of the transcriptome, which more often contains the mutations that affect tumor progression. WES can also be expanded to include untranslated regions and microRNA (miRNA)-binding sites. Because WES is faster and relatively inexpensive, it is often the best approach for clinically related research and patient studies (Koeppel et al. 2018).

Analysis of gene expression and transcriptome changes with WES or RNA-seq can aid in understanding tumor classification and help determine how specific groups of genes with mutations or altered expression levels can provide information of clinical utility such as tumor progression or response to therapies. Most tumors accumulate numerous genetic changes, but typically only a few genes have altered expression affecting cancer pathways that drive tumor progression. Targeted RNA-seq is a rapid and convenient method for obtaining expression data from a series of specific transcripts of interest related to a clinical phenotype, such as response to therapy or probability of disease recurrence.

Multigene panels are being increasingly used to provide precise genomic data to guide clinical decisions (Malone et al. 2020). One of the best examples of an expression panel used clinically is the Oncotype DX gene test designed to estimate the probability of recurrence for patients with a specific type of early breast cancer (Paik et al. 2004). The assay is performed using RNA extracted from formalin-fixed paraffin-embedded (FFPE) breast cancer tissue biopsies, using quantitative real-time reverse transcriptase quantitative polymerase chain reaction (RT-qPCR). The panel contains a set of five reference genes and 16 cancer-related genes. The probability of recurrence is derived from calculating the weighted expression of each of the 16 genes and classifying the chance of tumor recurrence as being low, intermediate, or high risk.

Chromosomal rearrangements that juxtapose two different genes together can form a fusion gene that encodes a fusion transcript, translated into a chimeric fusion oncoprotein. Many of the cancer gene fusions are strong driver mutations in neoplasia, and their identification can be helpful for diagnosis or is critical for effective treatment of some types of malignancy (Mertens et al. 2015). Older methods such as fluorescence in situ hybridization (FISH) and RT-qPCR have limited resolution and are low throughput. Targeted RNA-seq can simultaneously identify multiple fusion genes in a single tumor sample. This type of test enables better molecular classification into cancer subtypes and is also likely to increase the number of fusion genes in human cancer, including rare fusion genes and fusions of novel gene partners (Heyer et al. 2019).

13.3 Public Domain Transcriptomic Resources and Informatics

During the last years, sequencing technologies evolved rapidly. In the nineties sequencing, the human genome took almost 15 years; nowadays, next-generation sequencing can do the same analysis in a few hours. Since high-throughput sequencing technologies allow unlimited possibilities for analysis, there is a consensus that all genomic data should be available in the public domain (Conesa and Beck 2019). To this end, there are now a diversity of freely available online resources and practical algorithms and pipelines to work with multiomics and clinical data. Below we summarize some of the approaches currently being used for bioinformatics analysis in human cancer research.

The size and the complexity of the combined layers of genomic, transcriptional, and proteomic profiles available in the public domain pose a formidable challenge for new investigators in this area. Table 13.1 shows a summary of some of the excellent online resources and different types of analytical software to provide bioinformatics support for students and new researchers in this area to initiate projects using cancer genomic databases.

The Cancer Genome Atlas (TCGA) project has characterized over 20,000 primary cancer and matched normal samples for 33 cancer types (NIH, https://portal.gdc.cancer.gov/). TCGA currently has over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data available for public domain use. TCGA selected tumor type for study based on clinical needs in human cancer. Various practical algorithms were developed for analysis, and these shared datasets are easy to use and have good online support. For beginners, cBioPortal for Cancer Genomics is probably the best place to start investigating online transcriptomics. It is an open-access resource for interactive explorations of cancer genomics datasets from different origins such as TCGA, ICGC, Stand Up to Cancer (SU2C), and Memorial Sloan-Kettering Cancer Center (MSKCC). The International Cancer Genome Consortium (ICGC) is similar to a global initiative to build a comprehensive catalog of mutational abnormalities in the major tumor types (Zhang et al. 2019).

Both the ICGC and TCGA whole-genome sequencing studies recently formed a consortium to publish a meta-analysis of genomic features of representative tumor types (Campbell et al. 2020). RNA-sequencing data from TCGA and ICGC cohorts were collected and re-analyzed centrally for 1222 samples, including 1178 primary tumors, 67 metastases or local recurrences, and 153 matched normal tissue samples adjacent to the primary tumor. The data were uniformly processed to quantify normalized gene-level expression. The analysis can be used to detect splicing variation, allele-specific expression, fusion transcripts, alternative promoter usage, and any sites of RNA editing. The data generated by the consortium produced novel insights into the nature and timing of many mutational processes that shape somatic variation and impact cancer transcriptomics. Findings reported recently show generalized effects of somatic variants on transcription. In addition, the consortium study highlights the role of intratumoral heterogeneity on progression and the distinct evolutionary trajectory of each type of cancer (Calabrese et al. 2020).

Table 13.1 Informatics resources for transcriptomics analysis

Type	Informatics resource	Description	Host	Data	Website domain
Database	The Cancer Genome Atlas (TCGA)	TCGA is a major cancer genomics program that contains a molecular characterization of over 20,000 primary cancer.	NIH	WGS, WES, RNAseq	https://portal.gdc.cancer.gov
	The International Cancer Genome Consortium (ICGC)	ICGC is a similar comprehensive catalog of mutational abnormalities in the major tumor types. 22,000 tumors. ICGC/TCGA published combined data as the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium.	Global Alliance for Genomics & Health	WGS, WES, RNAseq	https://dcc.icgc.org/
	Genotypes and Phenotypes (dbGaP)	It is a database of data and results from studies that have investigated the interaction of genotype and phenotype in humans.	NCBI	Genotype, expression, epigenomic, genomic sequence, somatic mutation	https://www.ncbi.nlm.nih.gov/gap/
	LinkedOmics	It is a publicly available portal that includes multiomics data, from all 32 TCGA cancer types.	Zhang Lab	Multiomics	http://www.linkedomics.org/login.php
	cBioPortal for Cancer Genomics	It is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets.	MSKCC	WGS, WES, RNA-Seq, clinical data	https://www.cbioportal.org/
	Gene Expression Omnibus	It is an international public repository that archives and freely distribute high-throughput functional genomics data.	NCBI	microarray, NGS, high-throughput functional genomics	https://www.ncbi.nlm.nih.gov/geo/
	Tumor Immune Single-cell Hub (TISCH)	It is a scRNA-seq database that provides detailed cell-type annotation at the single-cell level.	CompGenomics	single-cell RNA seq	http://tisch.comp-genomics.org/

(continued)

Table 13.1 (continued)

	The Genotype-Tissue Expression (GTEx)	It is a comprehensive public resource to study tissue-specific gene expression and regulation. It encompasses 54 nondiseased tissue sites across nearly 1000 individuals.	Broad Institute of MIT and Harvard	WGS, WES, and RNA-Seq	https://gtexportal.org/home/
	NONCODE	It is a database dedicated to noncoding RNAs (excluding tRNAs and rRNAs) of 39 species.	Biologic Medicine Information Center of China	RNA-Seq	http://www.noncode.org/index.php
	National Cancer Institute Genomic Data Commons (NCI's GDC)	It is a database that supports the import and standardization of genomic and clinical data from cancer research programs.	NCI	WGS, WES, RNA-Seq, etc.	https://gdc.cancer.gov/
	The Protein Atlas	It is a program that aims to map all the human proteins in cells, tissues, and organs using an integration of various omics technologies	KTH Royal Institute of Technology	Multiomics	https://www.proteinatlas.org/
Pathway Analysis	Gene Set Enrichment Analysis	It is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states.	UC San Diego, Broad Institute	Molecular Signatures databases	https://www.gsea-msigdb.org/gsea/index.jsp
	The Database for Annotation, Visualization, and Integrated Discovery	It is a set of web-accessible programs of functional annotation tools to investigate the meaning of a large list of genes.	LHRI, The Frederick National Laboratory	Webtool	https://david.ncifcrf.gov/
	Cytoscape	It is an open source desktop application for large-scale network analysis and visualization with different plugins that enable addition of new features and functionalities according to the user's needs.	NIGMS, NRNB	Multiomics	https://cytoscape.org/

	Gene Ontology	It is the world's largest source of gene information; it has a connection with the Panther Classification system.	NHGRI	Multiomics	http://geneontology.org/
Software	Bioconductor Package for R	It is a comprehensive web archive of tools for the analysis and comprehension of high-throughput genomic data	Bioconductor	R packages	https://www.bioconductor.org/
	Orange Data Mining	It is a software that builds data analysis workflows visually, with a large, diverse toolbox.	University of Ljubljana	Multiple data types	https://orangedatamining.com/
	Galaxy	Galaxy is an open-source, web-based platform for data-intensive biomedical research.	Penn State, Johns Hopkins University, and Oregon Health & Science University	Multiomics	https://galaxyproject.org/
	Next-Generation Clustered Heat Map (NG-CHM) Viewer	It is a graphical environment for exploration of clustered or nonclustered heat map data in a web browser.	MD Anderson, Bioinformatics	Webtool	https://bioinformatics.mdanderson.org/public-software/ngchm/
	ReactomeFIViz	It is a free software designed to find paths and network patterns related to cancer and other diseases; it enables the enrichment of genes, analysis of RNA-seq, and also scRNA-seq; it is easy to use.	OICR (Ontario Institute for cancer research)	Webtool and local domain	https://reactome.org/tools/reactome-fiviz
	Kallisto	It is a complete and powerful tool for scRNA-seq and RNA-seq analysis, but it does not have a graphical interface.	Pachter Lab (Caltech)	Local domain tool	https://pachterlab.github.io/kallisto/

13.4 Transcriptomic Networks and Cancer Pathway Analysis

Determining which pathways are activated by gene expression alterations in transcriptomic data is of central importance for understanding cancer biology and possible clinical utility. The development of tools like The Database for Annotation, Visualization and Integrated Discovery (DAVID), Gene Ontology (GO), Gene Set Enrichment Analysis (GSEA), Ingenuity Pathway Analysis (IPA), and the Bioconductor packages for R helped focus on specific cancer pathways of therapeutic or diagnostic importance (Paczkowska et al. 2020).

Co-expression analysis is the principal tool to group genes of similar functions that impact common pathways of clinical importance. When the samples being studied also have linked clinical information on tumor stage and grade, therapies, and patient outcome, this type of analysis can be very informative. Collectively, this information is analyzed by integrative bioinformatics to define which driver pathways are related to the observed clinical outcomes. Together, these findings can suggest tumor phenotypes activated by the transcriptional networks that impact the hallmarks of cancer (Hanahan and Weinberg 2011).

Gene Expression Omnibus (GEO) is the central repository to find different kinds of datasets worldwide, but it may require great familiarity and computational power since only raw data is available from this resource. For data visualization, Next-Generation Clustered Heat Map (NG-CHM) Viewer is a comprehensive web-based graphical environment that may help. For expression enrichment analysis, The Database for Annotation, Visualization, and Integrated Discovery (DAVID) is the most accessible web-based resource to use, while Gene Set Enrichment Analysis (GSEA) may provide more complex results. For data processing, preliminary analysis, GALAXY, and Orange Data mining software are comprehensive tools that may be useful even for beginners. At the same time, the Bioconductor Package for R is the complete tool for reliable analysis of different types, but that requires a broader knowledge in R programming.

13.5 The Tumor Microenvironment, Immune Evasion, and Immunotherapy Response

Immune evasion is an important hallmark of cancer associated with failed response to immunotherapy (Hanahan and Weinberg 2011). The evasion phenotype is greatly facilitated by the tumor microenvironment (TME), which can have a suppressive effect on the immune system and also works as a protective barrier for cancer cells. The TME comprises multiple components: the extracellular matrix, surrounding stromal cells, infiltrating immune cells, and various signaling molecules, all of which can influence the competing pressures of immune response vs. tumor growth (Havel et al. 2019). During tumor growth, regulatory cellular processes are lost, and there is an accumulation of somatic genetic alterations. These genomic changes

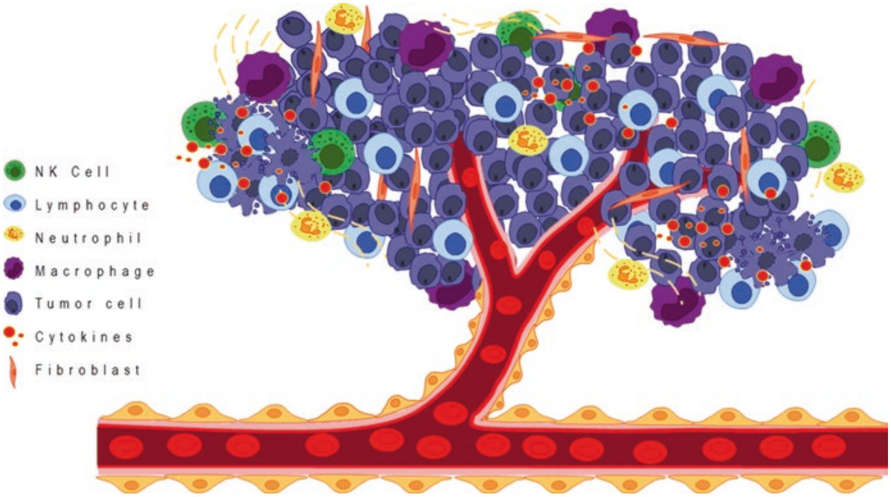


Fig. 13.1 Schematic diagram of tumor to iillustrate cellular heterogeniety in the microenvironment
Different cell phenotypes indicated on the left are depicted in different colors as rare cells intermingled with an excess of blue tumor cells. The central part of the tumor comprises blood vessels and vacularized regions on both sides

affect the tumor transcriptome, which can also influence the natural immune responses in the TME of developing cancer (see Fig. 13.1).

Immune checkpoint inhibition involves the use of specific agents that block the suppressive interactions between a developing tumor and the defensive immune system of the patient (Zhao et al. 2019). Among the checkpoint-blocking strategies, the two most prominent in terms of clinical success to date are the targeting of cytotoxic-T-lymphocyte-associated protein 4 (CTLA-4) and the interaction between programmed cell death 1 (PD-1) and PD-L1. Immune checkpoint proteins such as programmed PD-L1 downregulate the immune system and promote self-tolerance by suppressing T cell inflammatory activity against tumors so that blocking the expression of checkpoint proteins with immune checkpoint inhibitors usually restores the capacity of the immune system to recognize tumor cells and kill them (Koyama et al. 2016; Deng et al. 2018; Della Corte et al. 2019).

Immune checkpoint's normal function is to prevent autoimmunity and tissue damage during pathogenic infection. These molecules are inhibitory receptors expressed on the surfaces of T cells and tumor cells, and they mediate the functional interaction between these cells (Pardoll 2012). The process of adaptive immune resistance involves the engagement of immune checkpoint proteins on T cells by tumor cells to suppress T cells' cytotoxic capacity and enable tumor cells to escape immunosurveillance (Tumeh et al. 2014; Wu et al. 2014; Ribas 2015). T cell immune inhibition in response to cancer also involves the secretion of various inhibitory molecules such as cytokines and chemokines that decreases cytotoxic T lymphocyte function and limits the recruitment of anti-inflammatory cells, such as regulatory T cells and myeloid-derived suppressor cells and other immune cell types into the TME (Garcia et al. 2014; Kelderman et al. 2014; Yuan et al. 2016). There are now several computational approaches to estimating the nontumor cellular content of the TME, and these are described in the next section.

13.6 Computational Analysis of the TME from Bulk Tumor Transcriptome

Understanding how tumor intrinsic transcriptomic changes influence the immune cell content of the TME has been challenging because of the enormous excess of cancer cells in the tumor mass. In bulk extracted tumor tissue, the cancer transcriptome overwhelms the smaller proportion of gene expression data from immune and stromal cells in the TME. Recently there has been some success in predicting how the immune response changes in the TME using computational estimates of the abundances of member cell types in a mixed cell population based on analysis of transcriptomic data (Chen et al. 2018). These "immunoscores" use gene expression signatures of immune cell activity present in the tumor transcriptome to classify tumors into two broad types: immunologically nonresponsive or "cold" cancers and immunologically responsive or "hot" cancers (Maleki Vareki 2018). Immunologically cold tumors have a low mutation load, are immune tolerant against self-antigens, and lack infiltrating T cells (Yuan et al. 2016). In contrast, immunologically hot tumors have a variety of infiltrating T cells, which in turn reflects intrinsic T cell immune inhibition and extrinsic tumor-related T cell immunosuppression (Galon and Bruni 2019). Immunotherapy trials in "hot tumors" such as melanoma, urothelial, and lung cancer show that favorable responses are often observed. These tumors all have a pre-existing higher density of tumor-infiltrating lymphocytes (TILs) and expression of an interferon-associated gene signature (Gibney et al. 2016; Shien et al. 2016). In contrast, immunologically cold tumors such as pancreatic or prostate cancer have a lower TILs density in the TME and they cannot elicit a normal immune response to developing cancer. The genomics of tumors with such differing immune responses suggests various mutational changes can influence pathways of evasion and the tumor–immune interactions making a tumor immunologically cold (Thorsson et al. 2018).

Tumor-infiltrating immune cells can be quantified from RNA sequencing data of human tumors using various bioinformatics approaches (Finotello and Trajanoski 2018). In Table 13.2, we show examples of some of the recent computational methods that quantify immune cells from expression data of cell mixtures using marker genes coupled with GSEA or other scoring approaches that rely on deconvolution algorithms and immune cell expression signatures. ESTIMATE is one of the simplest methods of analyzing cellular heterogeneity based on transcriptomics, providing scores for tumor purity, immune infiltration, and stromal presence. CIBERSORTx

Table 13.2 Informatics resources to tumor microenvironment exploration using transcriptomics data

Tumor microenvironment resources	Description	Types of cells	Website domain
Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data (ESTIMATE)	ESTIMATE is an algorithm that provides researchers scores for tumor purity, the level of stromal cells present, and the infiltration level of immune cells in tumor tissues based on expression data in different platforms.	Not applicable	https://bioinformatics.mdanderson.org/estimate/
CIBERSORTx	CIBERSORTx is an analytical tool developed to impute gene expression profiles and provide an estimation of the abundances of member cell types in a mixed cell population, using gene expression data. CIBERSORTx allows users to process gene expression data representing a bulk admixture of different cell types or single-cell transcriptome sequencing.	22 immune cells	https://cibersortx.stanford.edu/index.php
xCELL	xCell is a web tool that performs cell type enrichment analysis from gene expression data for 64 immune and stromal cell types. xCell is a gene signatures-based method learned from thousands of pure cell types and applies a novel technique for reducing associations between closely related cell types.	64 immune and stromal cells	https://xcell.ucsf.edu/
Microenvironment Cell Populations-counter (MPC-counter)	MPC-counter is an algorithm that allows a robust quantification of the absolute abundance of eight immune and two stromal cell populations in heterogeneous tissues from transcriptomic data.	8 immune cells, endothelial cells, and fibroblasts	https://cit.ligue-cancer.net/mcp-counter/
TIminer	TIminer is an easy-to-use computational pipeline for mining tumor–immune cell interactions from next-generation sequencing data.	28 immune cells	https://icbi.i-med.ac.at/software/timiner/timiner.shtml

and xCELL are the easiest to use for a beginner in this area. CIBERSORTx generates a predictive analysis for 22 immune cells, whereas xCELL provides greater depth with estimates of 64 cell types, including immune cells, stromal cells, stem cells, and others. Figure 13.2a provides a schematic depiction of the analytical steps of most digital cytometry programs, such as CIBERSORT or xCell. These types of software typically use a signature matrix from clustering analyses and compare

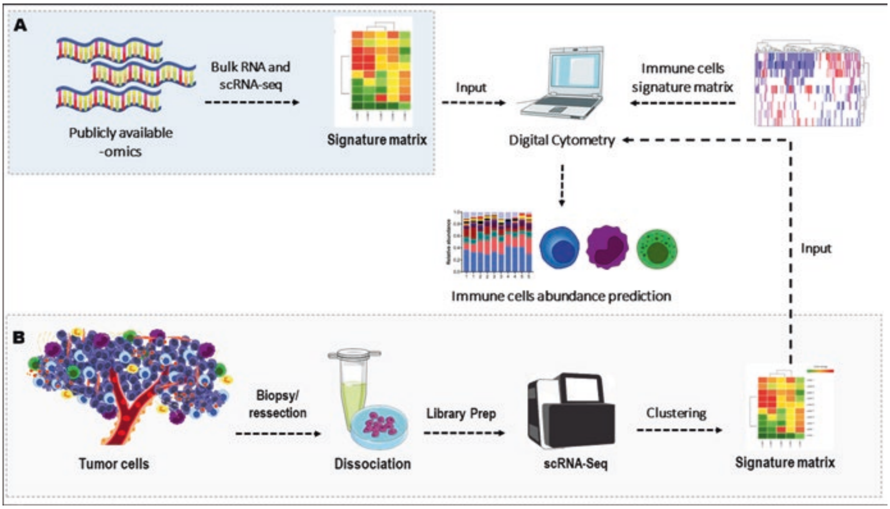


Fig. 13.2 Estimation of immune cell content in a tumor using (**a**) digital cytometry, and (**b**) scRNA-Seq methods
(**a**) Schematic representation of typical digital cytometric workflow. The pale blue shaded rectangle depicts publicly available sources of transcriptomic data used to generate gene expression signature matrices from tumors. These expression data are then computationally compared by digital cytometry algorithms to expression reference signatures of immune cells that are used to derive and estimate the relative abundances of the various immune cells in the tumor
(**b**) Schematic representation of scRNA-seq workflow. Schematic illustration of patient tumor to show presence of an excess of tumor cells mixed with heterogenous immune cells (see details in legend to Fig. 13.1). After a biopsy is removed from the tumor it is subject to tissue dissociation, sorting, and library preparation for high-throughput scRNA-seq. Each sample will comprise mixed cellular populations of tumor, immune and other rare nontumor cell-types. These expression data are then computationally compared by digital cytometry algorithms to expression reference signatures of tumor, immune, and noncancer cells to determine the relative abundances and single cell transcriptomics of each cell type sampled

expression levels to a previously inputted immune cell signature matrix. Thus, the immune cell abundance is displayed by the relative richness of the assigned transcripts related to each of the immune cells presented on each sample. All of the methods currently available can be used either in a web-based platform or as an R package.

As discussed below, new data from single-cell transcriptomics show that the computational cellular deconvolutions using bulk tumor transcriptomic data often fail to detect all rare cell types and subpopulations present in the TME (Yu et al. 2019). These observations draw attention to the need for experimental validation of predicted findings from digital cytometry using other methods.

13.7 Flow Cytometry and Image Analysis of the Tumor and TME

The classical methods for confirming the identities of nontumor cells and associated immune infiltrates in the TME are either by flow cytometry or microscopic image analysis. Both methods require specific antibody labeling strategies to identify immune cell subsets and their phenotypic features (Gerner et al. 2012; Bayne and Vonderheide 2013). These two strategies provide uniquely different information, as microscopy allows spatial appreciation of cellular subsets, whereas flow cytometry provides high throughput and broader quantification of cellular changes. Flow cytometry is one of the most widely used immune profiling techniques for characterizing the function of cells by exploring protein expression, cell subset frequency, cell function, immunophenotype, and ploidy (Maecker et al. 2012; Van Dongen et al. 2012; Streitz et al. 2013). Immunofluorescence image analysis is a complementary approach to flow cytometry, providing accurate information of immune cell types and their specific cell-to-cell interactions in the tumor microenvironment (Koelzer et al. 2019). Both flow cytometry and imaging approaches are invaluable for investigating the common nontumor cell types in the TME and understanding how their presence might influence immunotherapy response. This information is now supported by newer single-cell sequencing methods that are providing more detailed transcriptomic descriptions of the diversity of cell types in the TME.

13.8 Single-Cell mRNA Sequencing Analysis of the TME

Single-cell mRNA sequencing (scRNA-seq) enables researchers to distinguish the various cell types present in a dissociated tissue sample based on cellular gene expression levels. The method is typically performed for many hundreds to thousands of cells in a single experiment (Fig. 13.2b). Recent studies have shown that analyzing gene expression at the level of individual cells provides a much greater depth of analysis than earlier bulk methods (Lim et al. 2020). When combined with DNA sequencing, scRNA-seq allows appreciation of the *in vivo* impact of genomic alterations on gene expression. Importantly, scRNA-seq can assess the mutational variability and transcriptional pathways in cell populations present in the bulk tumor and the TME in an unbiased fashion at the level of individual cells. Gene expression analysis at the single-cell level can reliably distinguish neoplastic from nonneoplastic cells, to correlate paracrine-signaling pathways between neoplastic cells and the immune cells in the TME and surrounding stroma (Müller and Diaz 2017).

The TME is largely divided into the immune and stromal components, which can both be readily resolved using scRNA-seq methods. Investigations of the single-cell transcriptomics of the immune TME is an area of intense interest due to the growing use and success of immunotherapy in some tumor types, but apparent lack of response in other cancers. Recent studies have used scRNA-seq to profile T cells in tumors that fail to respond to treatment and have shown that a suppressive immune microenvironment is correlated with poor prognosis, in which increased T cell exhaustion signatures and decreased activated T cells were associated with clinical progression (Savas et al. 2018; Peng et al. 2019).

One of the most frequent immune cells present in the TME of solid tumors is the tumor-associated macrophage (TAM). TAMs have been previously classified by flow cytometric and imaging methods as M1 (inflammatory) or M2 (tumor-promoting). Analysis using scRNA-seq has shown that there is a continuum of macrophage transcriptomic programs, with a large diversity of cell states, suggesting that this traditional classification of TAMs may need refining (Azizi et al. 2018). Other scRNA-seq studies indicate that signaling from the TME may promote the differentiation of immature myeloid cells toward an immunosuppressive phenotype (Song et al. 2019). Single-cell analysis has also shown TAMs to be transcriptionally distinct from monocytes and their respective tissue-resident macrophages (Cassetta et al. 2019).

Most scRNA-seq studies published to date have centered on research applications and refinements of this powerful new technology. As scRNA-seq becomes more widely used the platform is likely to have more applications in cancer immunotherapy and in single-cell genomic classifications at earlier stages in the disease course (Lim et al. 2020).

13.9 Future Directions for Cancer Transcriptomics

Emerging data shows that a complete understanding of the dynamics of gene expression in cancer and the cross-talk between tumor and immune cells requires more detailed transcriptomic maps of tumor sections. In 2020, spatially resolved transcriptomics was designated method of the year in recognition of the potential of this technology in many areas of life science (Marx 2021). In the TME, there can be several subpopulations of cancer cells and intermingled nontumor cells that differ from each other completely in terms of both structural features and gene expression levels based on their location within the tumor. Proximity to endothelial cells, well-vascularized regions or areas of necrosis can all profoundly influence local gene expression. Spatially resolved transcriptomics provides a map of the spatial organization and the exact positions of variation in gene expression. The depth of spatial analysis is much greater than that obtained by bulk transcriptomic or scRNA-seq experiments (Asp et al. 2020). Spatial transcriptomic maps could be an important future tool for precision medicine in heterogeneous tumors where locally acting immune responses and clonal niches could be crucial for treatment decisions (Maniatis et al. 2021).

There is an increasing interest in using artificial intelligence (AI) to aid in the analysis of histological and other transcriptomic data (Yoosuf et al. 2020). AI typically uses deep neural networks to perform complex operations capable of capturing patterns or models that are not recognizable by traditional statistical methods

(Koelzer et al. 2019). Deep learning based on neural networks can also be used in a discriminative way to identify groups and subgroups of cells, to scale and improve the visual representation of scRNA-seq data to replace principal component analysis and classical unsupervised methods (Lin et al. 2017). Progress in recent years has shown that computing plays a crucial role in transcriptomics research, but it faces more challenges as new data is generated, leading to increased storage requirements and the need for supercomputers and life-science researchers highly skilled in AI and bioinformatics (Emani et al. 2021).

References

Asp M, Bergenstråhle J, Lundeberg J (2020) Spatially resolved transcriptomes—next generation tools for tissue exploration. BioEssays 42:1–16

Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, et al (2018) Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 174:1293–1308.e36

Bayne LJ, Vonderheide RH (2013) Multicolor flow cytometric analysis of immune cell subsets in tumor-bearing mice. Cold Spring Harb Protoc 2013:955–960

Calabrese C, Davidson NR, Demircioglu D, Fonseca NA, He Y, Kahles A, Van Lehmann K, Liu F, Shiraishi Y et al (2020) Genomic basis for RNA alterations in cancer. Nature 578:129–136

Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, Perry MD, Nahal-Bose HK, Ouellette BFF et al (2020) Pan-cancer analysis of whole genomes. Nature 578:82–93

Cassetta L, Baekkevold ES, Brandau S, Bujko A, Cassatella MA, Dorhoi A, Krieg C, Lin A, Loré K et al (2019) Deciphering myeloid-derived suppressor cells: isolation and markers in humans, mice and non-human primates. Cancer Immunol Immunother 68:687–697

Chen M, Wan L, Zhang J, Zhang J, Mendez L, Clohessy JG, Berry K, Victor J, Yin Q et al (2018) Deregulated PP1α phosphatase activity towards MAPK activation is antagonized by a tumor suppressive failsafe mechanism. Nat Commun 9:159

Cieślik M, Chinnaiyan AM (2018) Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet 19:93–109

Colombo PE, Milanezi F, Weigelt B, Reis-Filho JS (2011) Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. Breast Cancer Res 13:2–15

Conesa A, Beck S (2019) Making multi-omics data accessible to researchers. Sci Data 6:1–4

Della Corte CM, Barra G, Ciaramella V, Di Liello R, Vicidomini G, Zappavigna S, Luce A, Abate M, Fiorelli A et al (2019) Antitumor activity of dual blockade of PD-L1 and MEK in NSCLC patients derived three-dimensional spheroid cultures. J Exp Clin Cancer Res 38:1–12

Deng R, Fan F yi, Yi H, Liu F, He G cui, Sun H ping, Su Y (2018) PD-1 blockade potentially enhances adoptive cytotoxic T cell potency in a human acute myeloid leukaemia animal model. Hematology 23:740–746

Emani PS, Warrell J, Anticevic A, Bekiranov S, Gandal M, McConnell MJ, Sapiro G, Aspuru-Guzik A, Baker JT et al (2021) Quantum computing at the frontiers of biological sciences. Nat Methods. https://doi.org/10.1038/s41592-020-01004-3

Finotello F, Trajanoski Z (2018) Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother 67:1031–1040

Galon J, Bruni D (2019) Approaches to treat immune hot, altered and cold tumors with combination immunotherapies. Nat Rev Drug Discov 18:197–218

Garcia AJ, Ruscetti M, Arenzana TL, Tran LM, Bianci-Frias D, Sybert E, Priceman SJ, Wu L, Nelson PS et al (2014) Pten null prostate epithelium promotes localized myeloid-derived suppressor cell expansion and immune suppression during tumor initiation and progression. Mol Cell Biol 34:2017–2028

Gerner MY, Kastenmuller W, Ifrim I, Kabat J, Germain RN (2012) Histo-cytometry: a method for highly multiplex quantitative tissue imaging analysis applied to dendritic cell subset microanatomy in lymph nodes. Immunity 37:364–376

Gibney GT, Weiner LM, Atkins MB (2016) Predictive biomarkers for checkpoint inhibitor-based immunotherapy. Lancet Oncol 17:e542–e551

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: The next generation. Cell 144:646–674

Havel JJ, Chowell D, Chan TA (2019) The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer 19:133–150

Heyer EE, Deveson IW, Wooi D, Selinger CI, Lyons RJ, Hayes VM, O'Toole SA, Ballinger ML, Gill D et al (2019) Diagnosis of fusion genes using targeted RNA sequencing. Nat Commun 10:1–12

Kelderman S, Schumacher TNM, Haanen JBAG (2014) Acquired and intrinsic resistance in cancer immunotherapy. Mol Oncol 8:1132–1139

Koelzer VH, Sirinukunwattana K, Rittscher J, Mertz KD (2019) Precision immunoprofiling by image analysis and artificial intelligence. Virchows Arch 474:511–522

Koeppel F, Bobard A, Lefebvre C, Pedrero M, Deloger M, Boursin Y, Richon C, Chen-Min-Tao R, Robert G et al (2018) Added value of whole-exome and transcriptome sequencing for clinicalmolecular screenings of advanced cancer patients with solid tumors. Cancer J (United States) 24:153–162

Koyama S, Akbay EA, Li YY, Herter-Sprie GS, Buczkowski KA, Richards WG, Gandhi L, Redig AJ, Rodig SJ et al (2016) Adaptive resistance to therapeutic PD-1 blockade is associated with upregulation of alternative immune checkpoints. Nat Commun 7:1–9

Lim B, Lin Y, Navin N (2020) Advancing cancer research and medicine with single-cell genomics. Cancer Cell 37:456–470

Lin C, Jain S, Kim H, Bar-Joseph Z (2017) Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Res 45:1–11

Macgregor PF, Squire JA (2002) Application of Microarrays to the Analysis of Gene Expression in Cancer. Clin Chem 48:1170–1177

Maecker HT, McCoy JP, Nussenblatt R (2012) Standardizing immunophenotyping for the Human Immunology Project. Nat Rev Immunol 12:191–200

Maleki Vareki S (2018) High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors. J Immunother Cancer 6:4–8

Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL (2020) Molecular profiling for precision cancer therapies. Genome Med 12:1–19

Maniatis S, Petrescu J, Phatnani H (2021) Spatially resolved transcriptomics and its applications in cancer. Curr Opin Genet Dev 66:70–77

Marx V (2021) Method of the Year: spatially resolved transcriptomics. Nat Methods 18:9–14

Mertens F, Johansson B, Fioretos T, Mitelman F (2015) The emerging complexity of gene fusions in cancer. Nat Rev Cancer 15:371–381

Müller S, Diaz A (2017) Single-cell mRNA sequencing in cancer research: Integrating the genomic fingerprint. Front Genet 8:1–10

Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, Mee MW, Boutros PC, Abascal F et al (2020) Integrative pathway enrichment analysis of multivariate omics data. Nat Commun 11:1–16

Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817–2826

Pardoll DM (2012) The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer 12:252–264

Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, Liu L, Huang D, Jiang J et al (2019) Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res 29:725–738

Ribas A (2015) Adaptive immune resistance: How cancer protects from immune attack. Cancer Discov 5:915–919

Savas P, Virassamy B, Ye C, Salim A, Mintoff CP, Caramia F, Salgado R, Byrne DJ, Teo ZL et al (2018) Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat Med 24:986–993

Shien K, Papadimitrakopoulou VA, Wistuba II (2016) Predictive biomarkers of response to PD-1/PD-L1 immune checkpoint inhibitors in non–small cell lung cancer. Lung Cancer 99:79–87

Song Q, Hawkins GA, Wudel L, Chou PC, Forbes E, Pullikuth AK, Liu L, Jin G, Craddock L et al (2019) Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. Cancer Med 8:3072–3085

Streitz M, Miloud T, Kapinsky M, Reed MR, Magari R, Geissler EK, Hutchinson JA, Vogt K, Schlickeiser S et al (2013) Standardization of whole blood immune phenotype monitoring for clinical trials: panels and methods from the ONE study. Transplant Res 2:17

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, et al (2018) The immune landscape of cancer. Immunity 48:812–30.e14

Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJM, Robert L, Chmielowski B, Spasic M, Henry G et al (2014) PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature 515:568–571

Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, et al (2017) A pathology atlas of the human cancer transcriptome. Science 357(80-):1–11

Van Dongen JJM, Lhermitte L, Böttcher S, Almeida J, Van Der Velden VHJ, Flores-Montero J, Rawstron A, Asnafi V, Lécrevisse Q et al (2012) EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia 26:1908–1975

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wu YM, Cieślik M, Lonigro RJ, Vats P, Reimers MA, Cao X, Ning Y, Wang L, Kunju LP et al (2014) PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature 515:568–571

Yoosuf N, Navarro JF, Salmén F, Ståhl PL, Daub CO (2020) Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. Breast Cancer Res 22:1–10

Yu X, Chen YA, Conejo-Garcia JR, Chung CH, Wang X (2019) Estimation of immune cell content in tumor using single-cell RNA-seq reference data. BMC Cancer 19:1–11

Yuan J, Hegde PS, Clynes R, Foukas PG, Harari A, Kleen TO, Kvistborg P, Maccalli C, Maecker HT et al (2016) Novel technologies and emerging biomarkers for personalized cancer immunotherapy. J Immunother Cancer:4

Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, Stein LD, Ferretti V (2019) The international cancer genome consortium data portal. Nat Biotechnol 37:367–369

Zhao SG, Lehrer J, Chang SL, Das R, Erho N, Liu Y, Sjöström M, Den RB, Freedland SJ et al (2019) The immune landscape of prostate cancer and nomination of PD-L2 as a potential therapeutic target. J Natl Cancer Inst 111:301–310

Chapter 14
MicroRNAs in Cancer

Adriane F. Evangelista, Ana Julia A. de Freitas, Muriele B. Varuzza, Rhafaela L. Causin, Tatiana T. Komoto, and Marcia M. C. Marques

14.1 MicroRNAs: Characterization and Biogenesis

Due to a collection of studies on understanding the mechanisms of microRNA (miRNA) regulation, it was possible to verify that the dysregulation of miRNA expression can lead to the development and progression of numerous diseases in humans, such as cancer (Ardekani and Naeini 2010; Iorio and Croce 2012; Lages et al. 2012; Condrat et al. 2020).

Aberrant miRNA expression in cancer cells is mainly attributed to the localization of approximately 50% of miRNAs in fragile sites or regions, which, in turn, are more susceptible to genomic changes and associated with tumorigenesis (Calin et al. 2004; Marquardt et al. 2020). Another important factor that has been pointed out by the researchers in the field is the epigenetic changes that might induce aberrant miRNA expression (Fabbri et al. 2019; Barbieri and Kouzarides 2020), such as the methylation of miRNA genes (Huang et al. 2014; De Vuyst et al. 2015; Rogeri et al. 2018; Del Pino et al. 2019), as well as the activation, or even the inhibition of the biogenesis machinery of these small molecules (Lee et al. 2003; Chendrimada et al. 2005; O'Donnell et al. 2005). The first evidence of the involvement of miRNAs in cancer was based on the identification of the coding sequence of miR-15 and miR-16 in the 13q14 chromosomal region (Calin et al. 2002), and this region is frequently deleted in chronic lymphocytic leukemia tumors. Since then, the dysregulation of miRNA expression has been demonstrated in a range of tumors, such

A. F. Evangelista · A. J. A. de Freitas · M. B. Varuzza · R. L. Causin · T. T. Komoto
Molecular Oncology Research Center, Barretos Cancer Hospital, São Paulo, Brazil

M. M. C. Marques (✉)
Molecular Oncology Research Center, Barretos Cancer Hospital, São Paulo, Brazil

Barretos School of Health Sciences—FACISB, São Paulo, Brazil

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_14

as breast cancer (Loh et al. 2019), colorectal cancer (Zhu et al. 2020), lymphoma (Fernandez-Mercado et al. 2015), and cervical cancer (Causin et al. 2021).

In humans, it has been estimated that about 2588 miRNAs regulate more than 60% of human genes (Friedman et al. 2009; Shu et al. 2017), and thus, they have been shown to regulate key cellular processes, such as cell proliferation, DNA repair, differentiation, metabolism, and apoptosis (Forterre et al. 2020). Since the discovery of miRNAs in 1993, several studies have been conducted to understand the involvement of these molecules in normal physiological processes and the onset of a wide variety of diseases. Over 4700 different types of human miRNAs have been described so far, and this number is increasing rapidly (Griffiths-Jones 2004; miRBase 2021).

miRNAs are small noncoding RNAs (~19 to 24 nucleotides) that originate from precursor RNAs and are involved in the posttranscriptional regulation of coding genes (Lin and Gregory 2015). This regulation of gene expression occurs at the posttranscriptional level through interaction with the 3′ untranslated region (3′UTR) of messenger RNAs (mRNA) (Sevignani et al. 2006).

miRNA biogenesis is a controlled process (Calin and Croce 2006); however, it has been reported that miRNAs are dysregulated in cancer. These mechanisms can cause the loss of critical biological processes, such as proliferation, differentiation (Houbaviy et al. 2003), apoptosis (Cheng et al. 2005), epithelial–mesenchymal transition (Harquail et al. 2012), invasion, and migration (Armand-Labit and Pradines 2017).

The miRNA profile contributes to the molecular classification of tumors and can be associated with diagnosis, staging, progression, prognosis, and response to treatment (Calin and Croce 2006). Studies have shown that miRNAs are present in body

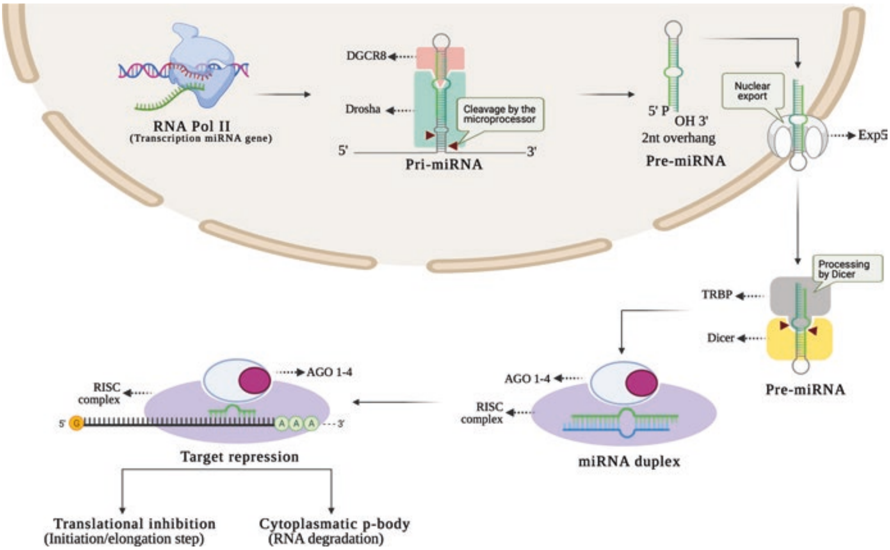


Fig. 14.1 The canonical pathway of miRNA biogenesis

fluids, which allow noninvasive identification, and suggest that serum miRNAs may act as potential biomarkers (Ono et al. 2015; Larrea et al. 2016; Polasik et al. 2017).

The canonical pathway of miRNA biogenesis (Fig. 14.1) (Kim et al. 2016) involves initial transcription by polymerase II (Pol II) of long RNAs called primary microRNA (pri-miRNA) (Bartel 2018), which possesses hundreds of nucleotides, and has 7-methyl guanosine at the 5′end, called cap 5′, and a polyadenylated 3′ tail (Cai et al. 2004; Lee et al. 2004; Neumeier and Meister 2020). A representative molecule of pri-miRNA exhibits a complex secondary structure, in the form of tweezers, each with a double-stranded rod and a handle containing noncomplementary bases and flank sequences (Bartel 2004). Then, the pri-miRNA is processed by a multiprotein complex called a multiprocessor. This process gives rise to a precursor RNA (pre-miRNA), which has approximately 70–120 nucleotides. The multiprocessor complex is composed of the RNase III enzyme Drosha (Lee et al. 2003), which binds to a cofactor, called gene 8 of the critical region of DiGeorge syndrome (DGCR8) or Pasha (Denli et al. 2004; Gregory et al. 2004; Landthaler et al. 2004).

The newly transcribed pre-miRNA is then exported to the cytoplasm via exportin 5 (Exp-5), which is a Ran-dependent nuclear transport receptor protein (Lund et al. 2004; Yi et al. 2003). In the cytoplasm, the pre-miRNA clamp is processed again into mature miRNA duplexes, with approximately 18–23 nucleotides. This process is mediated by another RNAase III enzyme, Dicer-1, which is functionally active when bound to the RNA-binding proteins in response to transactivation (TRBP) (Chendrimada et al. 2005). The strands of mature miRNA are then separated, which depends on several factors, such as the thermodynamic asymmetry of the duplex, and stability of the base pairs at the 5′ end. miRNAs mature, together with the RNA-binding proteins mentioned above and others, including the 6A gene containing trinucleotide repeat (TNRC6A), whose protein is associated with the catalytic Argonaut proteins 1-4 (AGO 1-4), giving rise to a microribonuclear protein complex (miRNP), called an RNA-induced silencing complex (RISC) (Schwarz et al. 2003). The miRNA strand with the most stable pairing generally acts as a guide tape, while the strands with stable base pairing are usually degraded (Okamura et al. 2009). This guide tape directs the RISC complex to the target mRNA through sequence complementarity and leads to its translational repression (Bartel 2018).

The mechanism of action of miRNAs involves the binding of miRNA to the 3′ UTR of the target mRNA, thereby resulting in the regulation of mRNA stability and protein synthesis (Bartel 2004, 2018). This posttranscriptional regulation mediated by miRNAs occurs through interaction (base pairing) in the 3′ UTR region and depends on the degree of complementarity between the miRNA and target mRNA. This interaction can lead to translation inhibition or mRNA degradation (Treiber et al. 2019). Imperfect pairing leads to the translation inhibition of the target mRNA, which is one of the major mechanisms of action of miRNAs in mammals (Friedman et al. 2009). The small size of miRNAs and their ability to function without complete base pairing means that a single miRNA can regulate multiple mRNA targets and that multiple miRNAs can cooperatively regulate the expression of a single mRNA (Lin and Gregory 2015).

Although the process of gene regulation through miRNA binding to the 3′ UTR region has been extensively studied, little is known about the regulation through the binding to the 5′ untranslated region (5′ UTR). This gene regulation mechanism of miRNAs has been demonstrated through several *in vitro*, *in vivo*, and *in silico* approaches, and a variety of tools have been developed to predict new targets and the function of miRNAs bound to the 5′ UTR (Da Sacco and Masotti 2012). Indeed, the gene regulation via the 5′ UTR is mainly known for its role in translational regulation, that is, dampening the translation through open upstream reading frames and secondary structures (Mignone et al. 2002). The first study to demonstrate such mechanism of the post-transcriptional regulation of gene expression by a 5′ UTRbound miRNA was conducted in *Drosophila melanogaster*. In this study, the authors reported that the interaction of miR-2 with the 5′ UTR region of the human *β*-globin gene inhibited the translation of the corresponding protein (Moretti et al. 2010).

Additionally, 5′ UTR regulation can provide a platform for miRNAs to bind and regulate the expression of target genes through a variety of other posttranscriptional mechanisms (Moretti et al. 2010; Dewing et al. 2012; Zhou and Rigoutsos 2014). It is important to note that genes that require fine regulation, such as growth factors, transcription factors, and proto-oncogenes, have been shown to possess longer 5′ UTR, providing more opportunity for such regulation (Mignone et al. 2002), as is the case with the regulation of *ATXN1* gene expression that occurs through the pairing of the 5′ UTR to miR-760 (Nitschke et al. 2020).

Classically, the interaction between the miRNA and mRNA targets leads to the translation repression of the mRNA (Ha and Kim 2014) through various mechanisms. Functionally, an miRNA can regulate the expression of protein coding as well as noncoding transcripts in a specific way, mainly through complementarity with the specific miRNA seed sequence (Bartel 2018). This interaction can impair the stability of mRNA and/or the translation of a specific protein, thereby leading to a reduction in the mRNA and/or protein expression levels (Filipowicz et al. 2008). The evidence on the mechanism of action of miRNAs has shown a new perspective on the complexity of posttranscriptional regulation exerted by these small molecules (Brennecke et al. 2005; Barbieri and Kouzarides 2020). Previously, the degree of complementarity between the miRNA and its mRNA targets was considered an important factor for determining the mode of translational repression (Bartel 2018). High complementarity can promote target cleavage mediated by Argonaute (AGO) protein, while complementarity to the seed sequence can lead to translation inhibition, and this is the most common mode of action of miRNAs (Moran et al. 2017). However, studies have revealed that approximately 60% of mammalian coding genes are regulated by miRNAs (Friedman et al. 2009; Lages et al. 2012). Although the regulatory mechanisms between the miRNAs and their respective targets are not yet fully understood; however, the literature has shown that these molecules are involved in numerous biological processes that are essential for cell survival (Iorio and Croce 2012; Lages et al. 2012) and that the imbalance of this regulation contributes to the development of numerous diseases (Ardekani and Naeini 2010; Tüfekci et al. 2014), especially cancer (Lages et al. 2012).

There are other noncanonical pathways of miRNA biogenesis, including a miRNA derived from an intronic region of mRNA by splicing independent of Drosha processing and a miRNA cleaved by specific poly (A) ribonuclease (PARN) independent of Dicer cleavage (Treiber et al. 2019).

Independent microRNAs (miRNAs) include mirtrons and tailed mirtrons, which are produced from the mRNA intronic regions during splicing. These RNAs are exported directly to the cytoplasm via exportin 1, without the need for Drosha cleavage. Dicer-independent miRNAs are processed from endogenous RNA transcripts by Drosha. Pre-miRNAs require AGO2 protein to complete their maturation within the cytoplasm because they are not long enough to act as a Dicer substrate (O'Brien et al. 2018).

In cancer, the mechanism of action of miRNAs is not different, given that these small molecules act in two different ways in the molecular pathways of tumor development. miRNAs are considered as "oncomiRs" when an increase in their expression contributes to the malignant transformation of normal cells, favoring the development and survival of these cells, or conversely, as tumor suppressor molecules, in reverse (Lages et al. 2012). The classification of miRNA between oncomiR or tumor suppressor miRNA is based on the evidence of its functional role in many types of tumors (Lages et al. 2012; Tutar et al. 2014). Several miRNAs have been described as tumor suppressor oncomirs (Jiang et al. 2009). In fact, miRNAs are natural cellular components and include an intrinsic signature that can guarantee specificity to the target. Thus the identification and understanding of the functional role of these unregulated molecules can be useful tools in the screening, diagnosis, and prognosis of several tumors, as well as in the development of personalized therapies. The miRNAs are natural cellular components possessing an intrinsic signature that guarantees target specificity. Therefore, the understanding of the functional role of these dysregulated molecules can act as useful tools for the screening, diagnosis, and prognosis of several tumors, as well as for the development of personalized cancer treatments.

14.2 Dysregulation of miRNA Expression in Human Cancer

miRNAs have been outstanding for molecular knowledge in human cancers, as they are involved in many cellular processes, as mentioned previously. Additionally, the small size of miRNAs and their ability to act without complete base pairing indicate that a single miRNA can regulate multiple mRNA targets and that multiple miRNAs can cooperatively regulate the expression of a single mRNA (Gebert and MacRae 2019). Recently, it has been proposed that the pathogenesis of cancer involves, among other macromolecules, the miRNAs, and their expression profiles are associated with the diagnosis, prognosis, and therapeutic responses of a variety of human cancers, as well as it can be considered as potential cancer biomarkers (O'Brien et al. 2018; Mollaei et al. 2019).

The miRNAs involved in neoplastic processes can be classified as oncomiRs and tumor suppressors, which means that the former acts as negatively regulating the tumor suppressor genes and the latter as an oncogene (Mollaei et al. 2019; Abd-Aziz et al. 2020). In summary, oncomiRs are the miRNAs that are related to tumor progression, as they are able to silence some tumor suppressor genes; therefore, they are mostly involved in processes, such as an increase in cell growth and metastasis. Conversely, miRNA tumor suppressors act inversely, which means inhibiting oncogenic genes related to tumorigenesis. Usually, this type of miRNAs are downregulated in tumors, and they are involved in several cellular mechanisms, such as genomic alterations, epigenetic alterations, and alterations in miRNA processing (Zhou et al. 2017; Mollaei et al. 2019).

For example, the group of polycistronic transcript derived miRNAs, known as the miR-17-92 cluster, which is located on chromosome 13q31, is found to be upregulated in the lung, colon, and gastric cancer, as well as lymphoma (Osada and Takahashi 2011; Concepcion et al. 2012; Fang et al. 2017), and it is considered to be potentially oncogenic because of its ability to modulate E2F1 expression and consequently inhibit apoptosis mediated by c-Myc via the p53 pathway (Rinaldi et al. 2007; Abd-Aziz et al. 2020). Moreover, this miRNA downregulates phosphatase, tensin homolog, and *RB2* (tumor suppressor genes) mediated by the protein kinase B signaling pathway to promote tumor cell survival (Shuang et al. 2013; Tan et al. 2018). Another example of oncomiR is miR-21, which is commonly found to be overexpressed in different tumors and has been demonstrated to have multiple targets and regulate different pathways and genes. MiR-21 has been shown to be involved in cell proliferation, metastasis, invasion, and chemoresistance (Abd-Aziz et al. 2020).

Conversely, the miRNA let-7, classified as a tumor suppressor, is downregulated in several tumor types, and its reduced expression is associated with poor prognosis (Boyerinas et al. 2010). Studies have demonstrated that the overexpression of let-7 can inhibit some important oncogenes related to tumor development and progression, such as *MYC*, *RAS*, *E2F1*, *E2F5*, *LIN28*, *ARID3B*, *HMGA2*, and long noncoding RNA *H19* (Chirshev et al. 2019). Additionally, let-7 expression correlates with the presence of cancer stem cells (CSCs). Therefore, once overexpressed, it can reduce the expression of CSC indicators, nestin and *CD133*, in glioblastomas, and that of *ALDH1* in breast cancer (Song et al. 2016; Sun et al. 2016; Chirshev et al. 2019). The miR-34 family (miR-34a, miR-34b, and miR-34c) is another important miRNA, similar to let-7, the expression of which is also reduced in many tumors, such as lung, breast, colon, and others (Li et al. 2013; Liu et al. 2011; Okada et al. 2014; Rokavec et al. 2014). They are described as miRNAs regulated by p53 (tumor suppressor), as they directly target the antiapoptotic proteins, Bcl-2 and SIRT1 (Li et al. 2013; Okada et al. 2014), and that their depletion is also related to metastasis and cancer recurrence, whereas their induction is associated with the improvement of apoptosis and efficacy of chemotherapy and radiation (Abd-Aziz et al. 2020).

Since more than 50% of miRNA genes are located in cancer-associated genomic sites (Friedman et al. 2009; Spengler et al. 2014), implying that miRNAs might play an important role in the pathogenesis of cancer. Therefore, previous studies have shown that miRNA expression can be used as a molecular biomarker (Tan et al. 2018); thus, based on their expression signatures, it is possible to differentiate normal cells from neoplastic cells at the molecular level and can be used to distinguish between several cancer types (Calin and Croce 2006).

An increasing number of studies on miRNAs have demonstrated that they play very important roles in the onset of cancer, such as proliferation, invasion, and metastasis (Abd-Aziz et al. 2020). The biggest problems related to cancer are associated with the angiogenesis and metastatic capacity of these cells. Therefore, the involvement of miRNAs in tumor metastasis has been intensely investigated in recent years. The role of miRNAs in metastasis was initially discovered by Ma et al. (2007), who demonstrated that miR-10b initiates invasion and metastasis in breast cancer. Additionally, some miRNAs are shown to modulate the expression of several genes related to both metastasis and angiogenesis; for example, miR-29c overexpression can downregulate the *VEGF* gene (vascular endothelial growth factor) and inhibit angiogenesis. Moreover, once miR-29c is upregulated in glioma cells, it has been shown to suppress migration and invasion of cells using an in vitro assay (Fan et al. 2013). Similarly, miR-497 has been shown to inhibit angiogenesis in breast cancer through targeting *VEGFR2* (Tu et al. 2015).

miRNAs are also associated with metalloproteinases (MMPs), which are essential for tissue remodeling in cancer angiogenesis and metastasis. The overexpression of miR-9 can inhibit *MMP14* levels, which leads to the reduction of angiogenesis, invasion, and metastasis in neuroblastoma cells (it has been proved using *in vitro* and *in vivo experiments*) (Zhang et al. 2012). Moreover, *MMP14* is a direct target of miR-181-5p in breast cancer cells, which may reduce their invasion, migration, and angiogenesis (Li et al. 2015; Lou et al. 2017). Together, these studies have revealed a balance between miRNAs as both the stimulators and inhibitors of metastasis, leading to the identification of several potential targets representing a molecular link between the loss of miRNA expression and specific behavior of a given tumor.

The other biggest problem associated with cancer is drug resistance; this process is complex and consists of many pathways, two of which are described here. First, the main mechanism of drug resistance is based on ATP-binding cassette proteins, which are a group of transmembrane proteins involved in the assimilation and secretion of cytotoxic compounds. P-glycoprotein (Pg-p) is responsible for drug resistance to a wide range of chemotherapeutic agents (Geretto et al. 2017). Therefore, many miRNAs regulate Pg-g expression and its activity (Garofalo and Croce 2013), such as miR-145 in intestinal cells (Ikemura et al. 2013), and miR-130 is related to cisplatin-resistant ovarian cancer cells (Yang et al. 2012). Conversely, miR-137 can reduce MCF-7 doxorubicin-resistance through targeting Y-box-binding protein-1 (YB-1) and subsequently downregulates the expression of Pg-p (Zhu et al. 2013). Another way by which tumor cells exhibit resistance is associated with DNA mismatch repair (MMR) genes. The lack of this pathway leads to drug resistance, inhibiting the cells to recognize the damage and then activate apoptosis. Additionally, indirect damage to this mechanism results in genome instability, and consequently an increase in the rate of mutation (Geretto et al. 2017).

While analyzing the information and assuming that miRNAs could be utilized as an alternate tool for cancer treatment, there are two therapeutic strategies aimed at re-establishing physiological miRNA expression in cancer cells, including the inhibition of oncomiR activity and restoration of tumor suppressor miRNA activity (Abd-Aziz et al. 2020). To inhibit the oncomiRs overexpressed in tumor cells, there are four strategies, including anti-miR oligonucleotides (AMO), locked nucleic acid (LNA), miRNA antagomiRs, and miRNA sponges (Shah et al. 2016). All these methods consist of miRNA inhibitors that exhibit complementarity to a single-stranded oligonucleotide and are able to isolate the endogenous miRNA in an unnatural structure, leading to the inactivation and elimination of mature miRNAs from the RISC complex (Shah et al. 2016; Abd-Aziz et al. 2020). Conversely, to restore miRNA, we commonly focus on inducing apoptosis or inhibiting cell tumor proliferation, which can be mediated using synthetic miRNA mimics or a viral vector expressing the miRNAs of interest (Shah et al. 2016). However, despite the attention that these strategies have received, some challenges remain to be addressed for their success. The main strategy is to effectively deliver either the miRNA antagonists or mimics directly into the tumor mass and to preserve their integrity and stability while in circulation (Yu et al. 2009; Jain and Stylianopoulos 2010; Paliwal et al. 2015). Another important point is the off-target side effects of these miRNAs (Wang et al. 2018; Segal and Slack 2020). Since they bind to multiple targets due to imperfect pairing in the 3′ UTR region, they exhibit the disadvantage of silencing many tumor suppressors, inducing potential toxicities, and reducing therapeutic effects (van Dongen et al. 2008; Meng and Lu 2017; Abd-Aziz et al. 2020). Therefore, gathering this knowledge, it can be inferred that many ongoing studies are still aiming to achieve success.

Finally, these findings are important not only because they represent a new field of research, but also because they finely dissect the molecular pathways in which miRNAs are involved in, such as tumor development. The abnormal expression of miRNAs in tumors, which is characterized by differential expression levels of the mature miRNA or miRNA precursor sequences compared to that in normal cells, has proven to be the main abnormality of the “miRNome” (the genome-wide set of miRNAs) observed in cancer cells.

14.3 Circulating miRNAs: Novel Biomarkers for Cancer

The potential use of miRNAs as biomarkers in several diseases has been explored, as these molecules are involved in important cellular processes, such as the regulation of posttranscriptional processes (Bracken et al. 2016). Considering that the dysregulation of miRNA expression is tissue specific, many studies have focused on exploring the potential use of miRNAs as biomarkers in cancer. These molecules could be used as tools for diagnosis and molecular subtyping, early detection, and therapeutics, and even to predict the disease (Wang et al. 2018; Sohel 2020).

Circulatory miRNAs are found in the circulation and can be assessed in biofluids, such as plasma, serum, urine, semen, and saliva, making it easier to examine a sample to be analyzed. The first tumor-specific miRNA was discovered in the serum of patients with B-cell lymphoma, wherein the increase in the expression of miR21 was associated with an increase in disease-free survival (Lawrie et al. 2008), and now, it has been identified in several types of cancer (Sohel 2020). Various studies have demonstrated the potential use of serum or plasma miRNAs as potential non-invasive biomarkers of different types of human cancers.

The potential use of miRNAs as biomarkers in the diagnosis and prognosis of cancer is primarily due to their stability and resistance for long periods under conditions that would normally lead to the degradation of other types of RNAs. However, the mechanism underlying miRNA resistance is not clearly understood yet. Moreover, two hypotheses have been proposed. The first is based on the molecular structure of the incorporating lipoprotein membrane-derived vesicles, such as high-density lipoproteins, exosomes, and microvesicles. The second hypothesis suggests that miRNAs are associated with protein complexes (Sohel 2020).

A study has shown that miRNAs are preserved in serum samples stored for 10 years (Valadi et al. 2007; Patnaik et al. 2010). The stability of these molecules can be partially explained by the discovery of these lipoprotein complexes, which are loaded with miRNAs (Valadi et al. 2007), messenger RNAs (El-Hefnawy et al. 2004), and proteins (Smalheiser 2007; Doyle and Wang 2019).

These vesicles are generally characterized into two major classes based on size: a smaller class of approximately 30–100 nm called exosomes, and a larger class of approximately 100 nm–1 μm called microvesicles (Doyle and Wang 2019). These microvesicles are formed by the internalization of the endosomal membrane to form multivesicular bodies that can subsequently merge with the plasma membrane, releasing the exosomes to the outside environment of cells (Théry et al. 2002; Michael et al. 2010). In circulation, these exosomes can export miRNAs to the recipient cells through endocytosis. After entering the cell, the delivered miRNAs are processed by the same machinery used for their biogenesis and can regulate gene expression in the recipient cells, leading to physiological modification.

Exosomes carrying miRNAs can be found not only in blood but also in other fluids, such as saliva and urine (Michael et al. 2010). Recently, exosomes have emerged as important mediators of cellular communication that are involved in normal physiological processes, such as immune response, lactation, and neuronal function (Admyre et al. 2007), as well as in the development and progression of diseases, such as cancer (Record 2013). In a recent study, the expression profile of miRNAs in the extracellular vesicles has been evaluated to predict the response upon treatment with anti-PD-1/PD-L1 in patients with non-small cell lung cancer (Shukuya et al. 2020).

In the context of cancer, this mechanism has been clearly demonstrated in glioblastoma patients where the tumor cells exported exosomes containing mRNA, miRNA, and angiogenic proteins that were detected through EGFRvIII receptors by normal cells, such as brain microvascular endothelial cells (Skog et al. 2008). In this study, it was shown that the cargoes delivered by the tumor-derived exosomes containing miRNAs could promote tumor progression. Furthermore, the results of this study showed that patients with cancer exhibited higher levels of exosomes in their plasma compared to those in control subjects. A recent review illustrated the application of exosomes as a glioblastoma biomarker (Giusti et al. 2017).

An interesting prospective study has shown that the tumor-specific miRNAs, such as miR-195, are differentially expressed in the circulatory system of women with breast cancer compared to those in the healthy control group. Furthermore, it was demonstrated that post tumor resection, the serum levels of miR-195 and let-7a were reduced (Heneghan et al. 2010). Another study showed that the analysis of the combined expression of miR-21, miR-210, miR-155, and miR-196a in plasma could help distinguish the patients with breast adenocarcinoma from the control subjects (Wang et al. 2009). The miRNA expression profile was determined using the entire genome of the serum sample of patients with triple-negative breast cancer, which revealed a signature of four miRNAs (miR-18b, miR-103, miR-107, and miR652) that could help predict tumor recurrence and overall survival (Kleivi Sahlberg et al. 2015). Moreover, the identification of miRNAs can provide more information regarding the molecular subtyping of these neoplasms (Souza et al. 2019).

The use of miRNAs as biomarkers has also been studied in other types of cancers other than breast cancer. A recent study identified a group of miRNAs (miR-19b3p, miR-26b-5p, miR-25-3p, and miR-301a-3p) in patients with ovarian cancer that target important genes involved in tumorigenesis, such as *PTEN*, *TP53*, and *ERBB2* (Penyige et al. 2019). In another study, the authors evaluated the expression of several miRNAs and identified a set of miRNAs (miR-873-3p, miR-149-5p, miR124-3p, miR-218-5p, miR-490-5p, miR-323a-3p, miR-10b-3p, miR-375, and miR-129-5p) with increased expression according to the advanced stage of neuroblastoma (Zeka et al. 2018). In prostate cancer, high expression of miR-17, mir-20a, mir-20b, and mir-106 has been shown to predict high-risk and high-stage diseases (Hoey et al. 2019). Several studies have identified the function of miRNAs and their potential use as biomarkers.

The use of miRNAs as cancer biomarkers relies on scientific evidence and the studies that aim to identify tissue-specific miRNAs detectable in biofluids to establish molecular signatures capable of characterizing the health status of the patients. Therefore, circulating miRNAs in body fluids and in extracellular compartments can act as hormones, which can trigger changes in the cellular gene expression through components secreted by a donor cell at the primary tumor site. Therefore, further studies are still required to be conducted to standardize the circulating miRNAs as biomarkers, as well as the techniques used to obtain high sensitivity and specificity (Sohel 2020; Valihrach et al. 2020).

14.4 Computational Approaches for miRNA Target Identification and Their Application in Cancer Research

Considering the role of the miRNAs in regulating gene expression of several cellular processes, the characterization of potential mRNA targets is crucial to understanding the mechanisms underlying the development of diseases, including cancer. (Bartel 2004; Sevignani et al. 2006; Peng and Croce 2016). It has been estimated that a single miRNAs can bind to approximately 200 transcripts on average (Friedman et al. 2009). A recent discovery of a new class of miRNA targets that seems not to depend on seed pairing suggests other complex rules for miRNA–mRNA interactions *in vivo* (Chipman and Pasquinelli 2019). However, the function of these molecules in cancer development and progression seems to be related to the regulation of a few critical targets, providing an entry point into the promising antisense miRNA-based therapeutic modalities (Peng and Croce 2016). Among the numerous challenges for the application of these strategies in clinical practice is the validation of their targets (Shah and Shah 2020; Marceca et al. 2021).

The first database designed to catalog the sequences of miRNAs as soon as they are identified was miRBase (www.mirbase.org). This database serves as a repository for sequences and annotations, providing access to virtually all published miRNAs (Griffiths-Jones 2004). Currently, the database is in version 22.1 (from October 2018) and consists of 38589 entries of 271 species, representing 1917 human miRNA precursor hairpins and 2654 mature sequences (Kozomara et al. 2019). This interface provides information about miRNA structure, genome location, and expression data and links out to some well-established third-party tools providing predicted and validated targets, such as TargetScan (Agarwal et al. 2015), DIANAmicroT (Maragkakis et al. 2011; Paraskevopoulou et al. 2013), miRDB (Wong and Wang 2015), TarBase (Karagkouni et al. 2018), and miRTarBase (Chou et al. 2018). The basic principles of the tools will be discussed below. It is important to emphasize that with the advance of large-scale technologies, especially next-generation sequencing, new miRNAs have been reported at a high rate (Cordero et al. 2012; Stäehler et al. 2012; Siddika and Heinemann 2021), and the miRBase increased more than a third over the previous release (Kozomara et al. 2019). Furthermore, other related databases reporting miRNA information are currently available, such as Rfam, which is synchronizing microRNA families with miRBase (Kalvari et al. 2021), the intragenic database miRIAD (Hinske et al. 2014), and the miRNA transcription start sites tracking program (mirSTP) (Liu et al. 2017).

Overall, the vast amount of available bioinformatics tools applied to miRNA research has been classified into different categories (Table 14.1). Initially, it was separated into two broader categories (Lindow 2011). The first includes those with precomputed predictions, in which the user does not need to perform all the steps but can search by miRNA name or related identification. The second consists of a server that allows the user to add their sequences for the analysis, making the prediction more versatile. This classification was mainly based on sequence-based

Table 14.1 General classification of microRNA tools according to selected publications

Classes	Description	Reference
Pre-computed predictions	Search by miRNA name or another identification;	Lindow (2011)
Web-based analysis tool	User sequence and parameter setting.	
Sequence and annotation	Sequence annotation registry databases;	Chen et al. (2019)
Target gene prediction	Characteristics of the miRNA sequence (structure-based, evolutionary conservation, machine learning, thermodynamic stability, integrated approach);	
Novel miRNA discovery	Next generation sequencing (NGS)-based;	
miRNA expression profiles	Next generation sequencing (NGS) and other experimental methods to identify miRNAs binding sites.	
Predictive methods (*de novo* predictions)	Characteristics of the miRNA sequence and/or based on the miRNA–mRNA interaction (seed pairing, thermodynamic stability, evolutionary conservation, accessibility target site, number of targets in the same 3′ UTR)	Riolo et al. (2020)
Predictive methods (Machine learning tools)	The miRNA target identification follows the miRNA–mRNA interactions with proven biological significance (pattern discrimination between actual and false targets);	
Operative strategy	Combination of tools derived from different predictions methods, allowing a good balance between sensitivity and specificity.	

target prediction algorithms, not considering the stand-alone platform tools and the more recent algorithms based on experimental methods (Lindow 2011). With the increasing number of tools for miRNAs analysis using several computational strategies, other classifications have emerged.

In this context, Chen et al. (2019) manually curated several reviews, accounting for more than 1000 miRNA bioinformatics tools published from 2003 to 2018 (last update) and organized in a comprehensive database called miRToolsGallery. This knowledge base classified the miRNA tools into four major categories: miRNA sequence and annotation, miRNA target gene prediction, novel miRNA discovery, and miRNA expression profiles. A more recent review categorized the miRNA tools available into three major categories: strategies for *de novo* predictions, machine learning tools, and operative strategy (Riolo et al. 2020). These classes are tentatives to organize the tools according to different criteria and are described in Table 14.1. The basic principles of some algorithms are discussed below.

This section mainly addresses the computational tools for target prediction and strategies that have been applied to the study of cancer. It is not an extensive list of methods available but a guide to the most common principles and an overview of the downstream analysis that can help users to extract useful biological information.

One of the basic principles widely used by these algorithms is seed pairing (Agarwal et al. 2015, 2018). In plants, complementary base pairing between miRNAs and their targets is almost perfect. Only a portion of the miRNA binds directly to the target in mammals, making prediction a challenging task (Bartel 2009). Thus, several rules have been established to identify targets based on sequence complementarity. This pairing is essential in positions 2–8 (seed region) at the 5′ end of the miRNA, or a high degree of similarity in the 3′ end of the miRNA can compensate for low complementarity in the seed region (Rajewsky 2006; Bartel 2009). Furthermore, the types of miRNA seed sequences vary in size and type, namely 8mer-A1 (positions 2–8 with match with an A opposite position 1), 7mer-m8 (position 2–8 match), 7mer-A1 (position 2–7 match with an A opposite position 1), 6mer (position 2–7 match), and offset-6mer (position 3–8 match) according to the sites matching in the miRNA seed region, with different regulatory efficiencies that are considered by these tools (Grimson et al. 2007; Friedman et al. 2009; Agarwal et al. 2015, 2018). Some additional criteria include the permission of G:U pairing, which may affect the repression capacity of the miRNA, among others.

Other well-used strategies include thermodynamic stability, accessibility to the target site, conservation patterns, among others. The miRNA target prediction tools based on thermodynamic stability consider the free energy estimate of the miRNA–mRNA interactions. When the free energy is low, more thermodynamically stable are the complexes (Akhtar et al. 2019). Furthermore, the mRNA 3′-UTR should be accessible for miRNA targeting and has been used by several tools. Algorithms based on 3′-UTR site accessibility usually rank the targets based on a calculated score (Robins et al. 2005). Moreover, several algorithms use conservation patterns for target prediction to reduce the number of false positives in such analyses. This approach is based on the principle that evolution selects and conserves useful biological functions (Riolo et al. 2020). However, not all sites are necessarily conserved, nor necessarily imply functionality. Nevertheless, conservation is of relevance in cancer. MicroRNAs reported to be oncogenes, or tumor suppressors are frequently conserved across species (Wang et al. 2010). Thus, considering the balance between the limitations and importance of such an approach in some cases, it is strongly recommended to combine these methods with other nonconservation models (Akhtar et al. 2019; Riolo et al. 2020).

As a few examples, the most popular computational tools that follow these rules and predict targets based mainly on sites in the 3′ UTR of the target are DIANAmicroT (Maragkakis et al. 2011; Paraskevopoulou et al. 2013), miRanda (Betel et al. 2010), PicTar (Krek et al. 2005), PITA (Kertesz et al. 2007), RNAhybrid (Rehmsmeier et al. 2004), and TargetScan (Grimson et al. 2007). Moreover, several tools have recently been developed to study the interactions between miRNAs and the 5′ UTR (or CDS) of target genes, such as miBridge, miRTar, miRWalk, and SfoldSTarMirDB (Da Sacco and Masotti 2012). Other widely used tools that allow the combination of several popular algorithms for target prediction are MAMI (MAMI 2021), miR-Gen (Megraw et al. 2007), and miRDip (Tokar et al. 2018).

Recently, new experimental methods for large-scale target validation have emerged, such as Stable Isotope Labeling by Amino acids in Cell culture (SILAC), which is a mass spectrometry (MS)-based quantitative proteomics used for miRNA target screening (Vinther et al. 2006), and Photoactivatable-Ribonucleoside-Enhanced

Table 14.2 Recommendations for selection of tools for microRNA target prediction and combination

Analysis steps	Reference
(1) Use of several algorithms with different methods for confirmation;	Witkos et al. (2011)
(2) Comparisons between mRNA and microRNA expression profiles;	
(3) Consideration of nearby sites that may act synergistically;	
(4) Experimental validation or subsequent functional assays.	
(1) Organism identification;	Kern et al. (2020)
(2) Selection of the target region (3′ UTR, 5′ UTR, coding sequences or entire mRNAs);	
(3) Input settings (sequences, expression levels, or both);	
(4) List type (target sites, target transcripts, or both);	
(5) If NGS data should be considered (if available);	
(6) Preference between the original features of the tools (seed sequence, free energy, site accessibility, machine learning) or a consensus of their prediction.	

Crosslinking and Immunoprecipitation (PAR-CLIP), which is a biochemical method used for identifying microRNA-containing ribonucleoprotein complexes (miRNPs) (Hafner et al. 2012), among others. These new methods have provided more robust results in this field and the creation of several computational tools. Other methods such as correlations, probabilistic methods, regression, and associations with transcription factors have been proposed to assess miRNA–mRNA networks based on gene expression data (Joung et al. 2007; Li et al. 2010).

Therefore, using a variety of tools is recommended to computationally search for the most representative targets and avoid false-positive data. Witkos et al. (2011) postulated that prediction methods considered efficient contain the four steps shown in Table 14.2. In line with this guide, a more recent approach described by Kern et al. (2020) suggests a six-step questionnaire that combines tools as a possible remedy for the limitations of using a distinct methodology (Table 14.2). An interactive webpage, freely available, can help users perform tool selection based on such criteria (https://ccb-web.cs.uni-saarland.de/mtguide). Both options are valuable for the best practices on miRNA target identification.

Furthermore, it is important to address the large amounts of data that are generated from large-scale transcriptome studies or large-scale target prediction. Usually, they can be summarized using functional enrichment analysis. The goal of this strategy is to provide a statistical method to estimate the enrichment, i.e., the higher-than-expected representation, of certain functional categories, excluding functional terms that could be identified by chance. Several tools use Fisher's exact test to estimate enrichment. Databases such as the popular DAVID (Database for Annotation, Visualization and Integrated Discovery) analyze data based on Gene Ontology (GO) functional categories and pathways from databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and others (Huang et al. 2009). Another option well used in cancer research includes the clusterProfiler package, which compares Gene Ontology and Disease Ontology terms, KEGG pathways, Network of Cancer Genes (NCG) information Molecular Signatures Database

(MSigDB), and customized ontology among gene clusters and has been updated more frequently (Yu et al. 2012).

Finally, there are some databases, which are generally manually curated from the literature, that compile information about diseases to derive biologically relevant information from lists of miRNAs and assist downstream analysis. Among the most well known are miR2Disease, which addresses 163 diseases (Jiang et al. 2009), and the Human microRNA Disease Database (HMDD) v3.2 (from March 2019), which provides information about microRNAs for 850 diseases and from approximately 30,000 experimentally supported miRNA–disease associations (Huang et al. 2019). In the cancer research, the ReactomeFIViz (Wu et al. 2014) provides information regarding pathway enrichment, cancer drugs, and the Cancer Gene Index (https://wiki.nci.nih.gov/display/cageneindex), and has been used recently by our group (Pessôa-Pereira et al. 2020; Evangelista et al. 2021). Furthermore, data from The Cancer Gene Atlas (TCGA) is an important resource for cancer research and databases, such as OMCD (OncomiR Cancer Database), has now supporting information from TCGA data and accounting for more than 9500 patients from 33 tumor types (Sarver et al. 2018). In summary, these tools provide evidence of the role of microRNAs from their targets, with approaches that can be used in cancer research.

References

Abd-Aziz N, Kamaruzman NI, Poh CL (2020) Development of MicroRNAs as potential therapeutics against cancer. J Oncol 2020:8029721. https://doi.org/10.1155/2020/8029721

Admyre C, Johansson SM, Qazi KR, Filén JJ, Lahesmaa R, Norman M, Neve EPA, Scheynius A, Gabrielsson S (2007) Exosomes with immune modulatory features are present in human breast milk. J Immunol 179(3):1969–1978. https://doi.org/10.4049/jimmunol.179.3.1969

Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. eLife 4:e05005. https://doi.org/10.7554/eLife.05005

Agarwal V, Subtelny AO, Thiru P, Ulitsky I, Bartel DP (2018) Predicting microRNA targeting efficacy in drosophila. Genome Biol 19(1):152. https://doi.org/10.1186/s13059-018-1504-3

Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD (2019) A practical guide to miRNA target prediction. Methods Mol Biol 1970:1–13. https://doi.org/10.1007/978-1-4939-9207-21

Ardekani AM, Naeini MM (2010) The role of MicroRNAs in human diseases. Avicenna J Med Biotechnol 2(4):161–179

Armand-Labit V, Pradines A (2017) Circulating cell-free microRNAs as clinical cancer biomarkers. Biomol Concepts 8(2):61–81. https://doi.org/10.1515/bmc-2017-0002

Barbieri I, Kouzarides T (2020) Role of RNA modifications in cancer. Nat Rev Cancer 20(6):303–322. https://doi.org/10.1038/s41568-020-0253-2

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116(2):281–297. https://doi.org/10.1016/s0092-8674(04)00045-5

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136(2):215–233. https://doi.org/10.1016/j.cell.2009.01.002

Bartel DP (2018) Metazoan MicroRNAs. Cell 173(1):20–51. https://doi.org/10.1016/j.cell.2018.03.006

Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 11(8):R90. https://doi.org/10.1186/gb-2010-11-8-r90

Boyerinas B, Park SM, Hau A, Murmann AE, Peter ME (2010) The role of let-7 in cell differentiation and cancer. Endocr Relat Cancer 17(1):F19–F36. https://doi.org/10.1677/ERC09-0184

Bracken CP, Scott HS, Goodall GJ (2016) A network-biology perspective of microRNA function and dysfunction in cancer. Nat Rev Genet 17(12):719–732. https://doi.org/10.1038/nrg.2016.134

Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of microRNA-target recognition. PLoS Biol 3(3):e85. https://doi.org/10.1371/journal.pbio.0030085

Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10(12):1957–1966. https://doi.org/10.1261/rna.7135204

Calin GA, Croce CM (2006) MicroRNA signatures in human cancers. Nat Rev Cancer 6(11):857–866. https://doi.org/10.1038/nrc1997

Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 99(24):15524–15529. https://doi.org/10.1073/pnas.242606799

Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A 101(9):2999–3004. https://doi.org/10.1073/pnas.0307323101

Causin RL, Freitas AJA, Trovo Hidalgo Filho CM, Reis RD, Reis RM, Marques MMC (2021) A systematic review of MicroRNAs involved in cervical cancer progression. Cells 10(3). https://doi.org/10.3390/cells10030668

Chen L, Heikkinen L, Wang C, Yang Y, Sun H, Wong G (2019) Trends in the development of miRNA bioinformatics tools. Brief Bioinformatics 20(5):1836–1852. https://doi.org/10.1093/bib/bby054

Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, Shiekhattar R (2005) TRBP recruits the dicer complex to ago2 for microRNA processing and gene silencing. Nature 436(7051):740–744. https://doi.org/10.1038/nature03868

Cheng AM, Byrom MW, Shelton J, Ford LP (2005) Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. Nucleic Acids Res 33(4):1290–1297. https://doi.org/10.1093/nar/gki200

Chipman LB, Pasquinelli AE (2019) MiRNA targeting – growing beyond the seed. Trends Genet 35(3):215–222. https://doi.org/10.1016/j.tig.2018.12.005, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7083087/

Chirshev E, Oberg KC, Ioffe YJ, Unternaehrer JJ (2019) Let-7 as biomarker, prognostic indicator, and therapy for precision medicine in cancer. Clin Transl Med 8(1):24. https://doi.org/10.1186/s40169-019-0240-y

Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, Chiew MY, Tai CS, Wei TY, Tsai TR, Huang HT, Wang CY, Wu HY, Ho SY, Chen PR, Chuang CH, Hsieh PJ, Wu YS, Chen WL, Li MJ, Wu YC, Huang XY, Ng FL, Buddhakosai W, Huang PC, Lan KC, Huang CY, Weng SL, Cheng YN, Liang C, Hsu WL, Huang HD (2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res 46:D296–D302. https://doi.org/10.1093/nar/gkx1067

Concepcion CP, Bonetti C, Ventura A (2012) The microRNA-17-92 family of microRNA clusters in development and disease. Cancer J 18(3):262–267. https://doi.org/10.1097/PPO.0b013e318258b60a

Condrat CE, Thompson DC, Barbu MG, Bugnar OL, Boboc A, Cretoiu D, Suciu N, Cretoiu SM, Voinea SC (2020) miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis. Cells 9(2). https://doi.org/10.3390/cells9020276

Cordero F, Beccuti M, Arigoni M, Donatelli S, Calogero RA (2012) Optimizing a massive parallel sequencing workflow for quantitative miRNA expression analysis. PLoS One 7(2):e31630. https://doi.org/10.1371/journal.pone.0031630

Da Sacco L, Masotti A (2012) Recent insights and novel bioinformatics tools to understand the role of microRNAs binding to 5′ untranslated region. Int J Mol Sci 14(1):480–495. https://doi.org/10.3390/ijms14010480

De Vuyst H, Franceschi S, Plummer M, Mugo NR, Sakr SR, Meijer CJLM, Heideman DAM, Tenet V, Snijders PJF, Hesselink AT, Chung MH (2015) Methylation levels of CADM1, MAL, and MIR124-2 in cervical scrapes for triage of HIV-infected, highrisk HPV-positive women in kenya. J Acquir Immune Defic Syndr 70(3):311–318. https://doi.org/10.1097/QAI.0000000000000744

Del Pino M, Sierra A, Marimon L, Martí Delgado C, Rodriguez-Trujillo A, Barnadas E, Saco A, Torné A, Ordi J (2019) CADM1, MAL, and miR124 promoter methylation as biomarkers of transforming cervical intrapithelial lesions. Int J Mol Sci 20(9). https://doi.org/10.3390/ijms20092262

Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the microprocessor complex. Nature 432(7014):231–235. https://doi.org/10.1038/nature03049

Dewing AST, Rueli RH, Robles MJ, Nguyen-Wu ED, Zeyda T, Berry MJ, Bellinger FP (2012) Expression and regulation of mouse selenoprotein p transcript variants differing in non-coding RNA. RNA Biol 9(11):1361–1369. https://doi.org/10.4161/rna.22290

Doyle LM, Wang MZ (2019) Overview of extracellular vesicles, their origin, composition, purpose, and methods for exosome isolation and analysis. Cells 8(7). https://doi.org/10.3390/cells8070727

El-Hefnawy T, Raja S, Kelly L, Bigbee WL, Kirkwood JM, Luketich JD, Godfrey TE (2004) Characterization of amplifiable, circulating RNA in plasma and its potential as a tool for cancer diagnostics. Clin Chem 50(3):564–573. https://doi.org/10.1373/clinchem.2003.028506

Evangelista AF, Oliveira RJ, Silva VAO, Vieira RADC, Reis RM, Marques MMC (2021) Integrated analysis of mRNA and miRNA profiles revealed the role of miR193 and miR-210 as potential regulatory biomarkers in different molecular subtypes of breast cancer. BMC Cancer 21(1):76. https://doi.org/10.1186/s12885-020-07731-2

Fabbri M, Girnita L, Varani G, Calin GA (2019) Decrypting noncoding RNA interactions, structures, and functional networks. Genome Res 29(9):1377–1388. https://doi.org/10.1101/gr.247239.118

Fan YC, Mei PJ, Chen C, Miao FA, Zhang H, Li Z (2013) MiR-29c inhibits glioma cell proliferation, migration, invasion and angiogenesis. J Neurooncol 115(2):179–188. https://doi.org/10.1007/s11060-013-1223-2

Fang LL, Wang XH, Sun BF, Zhang XD, Zhu XH, Yu ZJ, Luo H (2017) Expression, regulation and mechanism of action of the miR-17-92 cluster in tumor cells (review). Int J Mol Med 40(6):1624–1630. https://doi.org/10.3892/ijmm.2017.3164

Fernandez-Mercado M, Manterola L, Lawrie CH (2015) MicroRNAs in lymphoma: regulatory role and biomarker potential. Curr Genomics 16(5):349–358. https://doi.org/10.2174/1389202916666150707160147

Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of posttranscriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9(2):102–114. https://doi.org/10.1038/nrg2290

Forterre A, Komuro H, Aminova S, Harada M (2020) A comprehensive review of cancer MicroRNA therapeutic delivery strategies. Cancers (Basel) 12(7). https://doi.org/10.3390/cancers12071852

Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19(1):92–105. https://doi.org/10.1101/gr.082701.108

Garofalo M, Croce CM (2013) MicroRNAs as therapeutic targets in chemoresistance. Drug Resist Updat 16(3):47–59. https://doi.org/10.1016/j.drup.2013.05.001

Gebert LFR, MacRae IJ (2019) Regulation of microRNA function in animals. Nat Rev Mol Cell Biol 20(1):21–37. https://doi.org/10.1038/s41580-018-0045-7

Geretto M, Pulliero A, Rosano C, Zhabayeva D, Bersimbaev R, Izzotti A (2017) Resistance to cancer chemotherapeutic drugs is determined by pivotal microRNA regulators. Am J Cancer Res 7(6):1350–1371

Giusti I, Di Francesco M, Dolo V (2017) Extracellular vesicles in glioblastoma: role in biological processes and in therapeutic applications. Curr Cancer Drug Targets 17(3):221–235. https://doi.org/10.2174/1568009616666160813182959

Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R (2004) The microprocessor complex mediates the genesis of microRNAs. Nature 432(7014):235–240. https://doi.org/10.1038/nature03120

Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32(90001):109D–111D. https://doi.org/10.1093/nar/gkh023

Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27(1):91–105. https://doi.org/10.1016/j.molcel.2007.06.017

Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15(8):509–524. https://doi.org/10.1038/nrm3838

Hafner M, Lianoglou S, Tuschl T, Betel D (2012) Genome-wide identification of miRNA targets by PAR-CLIP. Methods 58(2):94–105. https://doi.org/10.1016/j.ymeth.2012.08.006

Harquail J, Benzina S, Robichaud GA (2012) MicroRNAs and breast cancer malignancy: an overview of miRNA-regulated cancer processes leading to metastasis. Cancer Biomark 11(6):269–280. https://doi.org/10.3233/CBM-120291

Heneghan HM, Miller N, Lowery AJ, Sweeney KJ, Newell J, Kerin MJ (2010) Circulating microRNAs as novel minimally invasive biomarkers for breast cancer. Ann Surg 251(3):499–505. https://doi.org/10.1097/SLA.0b013e3181cc939f

Hinske LC, França GS, Torres HAM, Ohara DT, Lopes-Ramos CM, Heyn J, Reis LFL, Ohno-Machado L, Kreth S, Galante PAF (2014) miRIAD—integrating microRNA inter- and intragenic data. Database 2014. https://doi.org/10.1093/database/bau099

Hoey C, Ahmed M, Fotouhi Ghiam A, Vesprini D, Huang X, Commisso K, Commisso A, Ray J, Fokas E, Loblaw DA, He HH, Liu SK (2019) Circulating miRNAs as noninvasive biomarkers to predict aggressive prostate cancer after radical prostatectomy. J Transl Med 17(1):173. https://doi.org/10.1186/s12967-019-1920-5

Houbaviy HB, Murray MF, Sharp PA (2003) Embryonic stem cell-specific MicroRNAs. Dev Cell 5(2):351–358. https://doi.org/10.1016/s1534-5807(03)00227-2

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57. https://doi.org/10.1038/nprot.2008.211

Huang YW, Kuo CT, Chen JH, Goodfellow PJ, Huang THM, Rader JS, Uyar DS (2014) Hypermethylation of miR-203 in endometrial carcinomas. Gynecol Oncol 133(2):340–345. https://doi.org/10.1016/j.ygyno.2014.02.009

Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q (2019) HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res 47:D1013–D1017. https://doi.org/10.1093/nar/gky1010

Ikemura K, Yamamoto M, Miyazaki S, Mizutani H, Iwamoto T, Okuda M (2013) MicroRNA-145 post-transcriptionally regulates the expression and function of pglycoprotein in intestinal epithelial cells. Mol Pharmacol 83(2):399–405. https://doi.org/10.1124/mol.112.081844

Iorio MV, Croce CM (2012) Causes and consequences of microRNA dysregulation. Cancer J 18(3):215–222. https://doi.org/10.1097/PPO.0b013e318250c001

Jain RK, Stylianopoulos T (2010) Delivering nanomedicine to solid tumors. Nat Rev Clin Oncol 7(11):653–664. https://doi.org/10.1038/nrclinonc.2010.139

Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y (2009) miR2disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 37:D98–D104. https://doi.org/10.1093/nar/gkn714

Joung JG, Hwang KB, Nam JW, Kim SJ, Zhang BT (2007) Discovery of microRNA-mRNA modules via population-based probabilistic learning. Bioinformatics 23(9):1141–1147. https://doi.org/10.1093/bioinformatics/btm045

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, Rivas E, Eddy SR, Finn RD, Bateman A, Petrov AI (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res 49:D192–D200. https://doi.org/10.1093/nar/gkaa1047

Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G, Vergoulis T, Dalamagas T, Hatzigeorgiou AG (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic Acids Res 46:D239–D245. https://doi.org/10.1093/nar/gkx1141

Kern F, Backes C, Hirsch P, Fehlmann T, Hart M, Meese E, Keller A (2020) What's the target: understanding two decades of in silico microRNA-target prediction. Brief Bioinformatics 21(6):1999–2010. https://doi.org/10.1093/bib/bbz111

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39(10):1278–1284. https://doi.org/10.1038/ng2135

Kim YK, Kim B, Kim VN (2016) Re-evaluation of the roles of DROSHA, export in 5, and DICER in microRNA biogenesis. Proc Natl Acad Sci U S A 113(13):E1881–E1889. https://doi.org/10.1073/pnas.1602532113

Kleivi Sahlberg K, Bottai G, Naume B, Burwinkel B, Calin GA, Børresen-Dale AL, Santarpia L (2015) A serum microRNA signature predicts tumor relapse and survival in triple-negative breast cancer patients. Clin Cancer Res 21(5):1207–1214. https://doi.org/10.1158/1078-0432.CCR-14-2011

Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. Nucleic Acids Res 47:D155–D162. https://doi.org/10.1093/nar/gky1141

Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N (2005) Combinatorial microRNA target predictions. Nat Genet 37(5):495–500. https://doi.org/10.1038/ng1536

Lages E, Ipas H, Guttin A, Nesr H, Berger F, Issartel JP (2012) MicroRNAs: molecular features and role in cancer. Front Biosci (Landmark Ed) 17:2508–2540. https://doi.org/10.2741/4068

Landthaler M, Yalcin A, Tuschl T (2004) The human DiGeorge syndrome critical region gene 8 and its D. melanogaster homolog are required for miRNA biogenesis. Curr Biol 14(23):2162–2167. https://doi.org/10.1016/j.cub.2004.11.001

Larrea E, Sole C, Manterola L, Goicoechea I, Armesto M, Arestin M, Caffarel MM, Araujo AM, Araiz M, Fernandez-Mercado M, Lawrie CH (2016) New concepts in cancer biomarkers: circulating miRNAs in liquid biopsies. Int J Mol Sci 17(5). https://doi.org/10.3390/ijms17050627

Lawrie CH, Gal S, Dunlop HM, Pushkaran B, Liggins AP, Pulford K, Banham AH, Pezzella F, Boultwood J, Wainscoat JS, Hatton CSR, Harris AL (2008) Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large b-cell lymphoma. Br J Haematol 141(5):672–675. https://doi.org/10.1111/j.13652141.2008.07077.x

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, Kim VN (2003) The nuclear RNase III drosha initiates microRNA processing. Nature 425(6956):415–419. https://doi.org/10.1038/nature01957

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23(20):4051–4060. https://doi.org/10.1038/sj.emboj.7600385

Li L, Xu J, Yang D, Tan X, Wang H (2010) Computational approaches for microRNA studies: a review. Mamm Genome 21(1):1–12. https://doi.org/10.1007/s00335-009-9241-2

Li L, Yuan L, Luo J, Gao J, Guo J, Xie X (2013) MiR-34a inhibits proliferation and migration of breast cancer through down-regulation of bcl-2 and SIRT1. Clin Exp Med 13(2):109–117. https://doi.org/10.1007/s10238-012-0186-5

Li Y, Kuscu C, Banach A, Zhang Q, Pulkoski-Gross A, Kim D, Liu J, Roth E, Li E, Shroyer KR, Denoya PI, Zhu X, Chen L, Cao J (2015) miR-181a-5p inhibits cancer cell migration and angiogenesis via downregulation of matrix metalloproteinase-14. Cancer Res 75(13):2674–2685. https://doi.org/10.1158/0008-5472.CAN-14-2875

Lin S, Gregory RI (2015) MicroRNA biogenesis pathways in cancer. Nat Rev Cancer 15(6):321–333. https://doi.org/10.1038/nrc3932

Lindow M (2011) Prediction of targets for microRNAs. Methods Mol Biol 703:311–317. https://doi.org/10.1007/978-1-59745-248-9_21

Liu C, Kelnar K, Liu B, Chen X, Calhoun-Davis T, Li H, Patrawala L, Yan H, Jeter C, Honorio S, Wiggins JF, Bader AG, Fagin R, Brown D, Tang DG (2011) The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. Nat Med 17(2):211–215. https://doi.org/10.1038/nm.2284

Liu Q, Wang J, Zhao Y, Li CI, Stengel KR, Acharya P, Johnston G, Hiebert SW, Shyr Y (2017) Identification of active miRNA promoters from nuclear run-on RNA sequencing. Nucleic Acids Res 45(13):e121–e121. https://doi.org/10.1093/nar/gkx318

Loh HY, Norman BP, Lai KS, Rahman NMANA, Alitheen NBM, Osman MA (2019) The regulatory role of MicroRNAs in breast cancer. Int J Mol Sci 20(19):DOI 10.3390/ijms20194940

Lou W, Liu J, Gao Y, Zhong G, Chen D, Shen J, Bao C, Xu L, Pan J, Cheng J, Ding B, Fan W (2017) MicroRNAs in cancer metastasis and angiogenesis. Oncotarget 8(70):115787–115802. https://doi.org/10.18632/oncotarget.23115

Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. Science 303(5654):95–98. https://doi.org/10.1126/science.1090599

Ma L, Teruya-Feldstein J, Weinberg RA (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature 449(7163):682–688. https://doi.org/10.1038/nature06174

MAMI (2021) MAMI microRNA meta-predictor | main. URL https://mami.med.harvard.edu/

Maragkakis M, Vergoulis T, Alexiou P, Reczko M, Plomaritou K, Gousis M, Kourtis K, Koziris N, Dalamagas T, Hatzigeorgiou AG (2011) DIANA-microT web server upgrade supports fly and worm miRNA target prediction and bibliographic miRNA to disease association. Nucleic Acids Res 39:W145–W148. https://doi.org/10.1093/nar/gkr294

Marceca GP, Tomasello L, Distefano R, Acunzo M, Croce CM, Nigita G (2021) Detecting and characterizing a-to-i microRNA editing in cancer. Cancers (Basel) 13(7):DOI 10.3390/cancers13071699

Marquardt S, Richter C, Pützer BM, Logotheti S (2020) MiRNAs targeting double strand DNA repair pathways lurk in genomically unstable rare fragile sites and determine cancer outcomes. Cancers (Basel) 12(4):DOI 10.3390/cancers12040876

Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG (2007) miRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res 35:D149–D155. https://doi.org/10.1093/nar/gkl904

Meng Z, Lu M (2017) RNA interference-induced innate immunity, off-target effect, or immune adjuvant? Front Immunol 8:331. https://doi.org/10.3389/fimmu.2017.00331

Michael A, Bajracharya SD, Yuen PST, Zhou H, Star RA, Illei GG, Alevizos I (2010) Exosomes from human saliva as a source of microRNA biomarkers. Oral Dis 16(1):34–38. https://doi.org/10.1111/j.1601-0825.2009.01604.x

Mignone F, Gissi C, Liuni S, Pesole G (2002) Untranslated regions of mRNAs. Genome Biol 3(3):REVIEWS0004. https://doi.org/10.1186/gb-2002-3-3-reviews0004

miRBase (2021) miRBase: the microRNA database. URL http://www.mirbase.org/

Mollaei H, Safaralizadeh R, Rostami Z (2019) MicroRNA replacement therapy in cancer. J Cell Physiol 234(8):12369–12384. https://doi.org/10.1002/jcp.28058

Moran Y, Agron M, Praher D, Technau U (2017) The evolutionary origin of plant and animal microRNAs. Nat Ecol Evol 1(3):27. https://doi.org/10.1038/s41559-016-0027

Moretti F, Thermann R, Hentze MW (2010) Mechanism of translational regulation by miR-2 from sites in the 5′ untranslated region or the open reading frame. RNA 16(12):2493–2502. https://doi.org/10.1261/rna.2384610

Neumeier J, Meister G (2020) siRNA specificity: RNAi mechanisms and strategies to reduce off-target effects. Front Plant Sci 11:526455. https://doi.org/10.3389/fpls.2020.526455

Nitschke L, Tewari A, Coffin SL, Xhako E, Pang K, Gennarino VA, Johnson JL, Blanco FA, Liu Z, Zoghbi HY (2020) miR760 regulates ATXN1 levels via interaction with its 5′ untranslated region. Genes Dev 34(17):1147–1160. https://doi.org/10.1101/gad.339317.120

O'Brien J, Hayder H, Zayed Y, Peng C (2018) Overview of MicroRNA biogenesis, mechanisms of actions, and circulation. Front Endocrinol (Lausanne) 9:402. https://doi.org/10.3389/fendo.2018.00402

O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT (2005) c-mycregulated microRNAs modulate e2f1 expression. Nature 435(7043):839–843. https://doi.org/10.1038/nature03677

Okada N, Lin CP, Ribeiro MC, Biton A, Lai G, He X, Bu P, Vogel H, Jablons DM, Keller AC, Wilkinson JE, He B, Speed TP, He L (2014) A positive feedback between p53 and miR-34 miRNAs mediates tumor suppression. Genes Dev 28(5):438–450. https://doi.org/10.1101/gad.233585.113

Okamura K, Liu N, Lai EC (2009) Distinct mechanisms for microRNA strand selection by drosophila argonautes. Mol Cell 36(3):431–444. https://doi.org/10.1016/j.molcel.2009.09.027

Ono S, Lam S, Nagahara M, Hoon DSB (2015) Circulating microRNA biomarkers as liquid biopsy for cancer patients: Pros and cons of current assays. J Clin Med 4(10):1890–1907. https://doi.org/10.3390/jcm4101890

Osada H, Takahashi T (2011) let-7 and miR-17-92: small-sized major players in lung cancer development. Cancer Sci 102(1):9–17. https://doi.org/10.1111/j.13497006.2010.01707.x

Paliwal SR, Paliwal R, Vyas SP (2015) A review of mechanistic insight and application of pH-sensitive liposomes in drug delivery. Drug Deliv 22(3):231–242. https://doi.org/10.3109/10717544.2014.882469

Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res 41:W169–W173. https://doi.org/10.1093/nar/gkt393

Patnaik SK, Mallick R, Yendamuri S (2010) Detection of microRNAs in dried serum blots. Anal Biochem 407(1):147–149. https://doi.org/10.1016/j.ab.2010.08.004

Peng Y, Croce CM (2016) The role of MicroRNAs in human cancer. Signal Transduct Target Ther 1(1):1–9. https://doi.org/10.1038/sigtrans.2015.4, URL https://www.nature.com/articles/sigtrans20154

Penyige A, Márton E, Soltész B, Szilágyi-Bónizs M, Póka R, Lukács J, Széles L, Nagy B (2019) Circulating miRNA profiling in plasma samples of ovarian cancer patients. Int J Mol Sci 20(18). https://doi.org/10.3390/ijms20184533

Pessôa-Pereira D, Evangelista AF, Causin RL, da Costa Vieira RA, Abrahão-Machado LF, Santana IVV, da Silva VD, de Souza KCB, de Oliveira-Silva RJ, Fernandes GC, Reis RM, Palmero EI, Marques MMC (2020) miRNA expression profiling of hereditary breast tumors from BRCA1- and BRCA2-germline mutation carriers in brazil. BMC Cancer 20(1):143. https://doi.org/10.1186/s12885-020-6640-y

Polasik A, Tzschaschel M, Schochter F, de Gregorio A, Friedl TWP, Rack B, Hartkopf A, Fasching PA, Schneeweiss A, Müller V, Huober J, Janni W, Fehm T (2017) Circulating tumour cells, circulating tumour DNA and circulating MicroRNA in metastatic breast carcinoma – what is the role of liquid biopsy in breast cancer? Geburtshilfe Frauenheilkd 77(12):1291–1298. https://doi.org/10.1055/s-0043-122884

Rajewsky N (2006) microRNA target predictions in animals. Nat Genet 38(Suppl):S8–S13. https://doi.org/10.1038/ng1798

Record M (2013) Emerging concepts of tumor exosome–mediated cell-cell communication. In: Zhang HG (ed) Emerging concepts of tumor exosome–mediated cell-cell communication, 1st edn. Springer, New York, pp 47–68. https://doi.org/10.1007/978-14614-3697-3, URL https://www.springer.com/gp/book/9781461436966

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10(10):1507–1517. https://doi.org/10.1261/rna.5248604

Rinaldi A, Poretti G, Kwee I, Zucca E, Catapano CV, Tibiletti MG, Bertoni F (2007) Concomitant MYC and microRNA cluster miR-17-92 (c13orf25) amplification in human mantle cell lymphoma. Leuk Lymphoma 48(2):410–412. https://doi.org/10.1080/10428190601059738

Riolo G, Cantara S, Marzocchi C, Ricci C (2020) miRNA targets: from prediction tools to experimental validation. Methods Protoc 4(1). https://doi.org/10.3390/mps4010001

Robins H, Li Y, Padgett RW (2005) Incorporating structure to predict microRNA targets. PNAS 102(11):4006–4009. https://doi.org/10.1073/pnas.0500775102, URL https://www.pnas.org/content/102/11/4006

Rogeri CD, Silveira HCS, Causin RL, Villa LL, Stein MD, de Carvalho AC, Arantes LMRB, Scapulatempo-Neto C, Possati-Resende JC, Antoniazzi M, Longatto-Filho A, Fregnani JHTG (2018) Methylation of the hsa-miR-124, SOX1, TERT, and LMX1a genes as biomarkers for precursor lesions in cervical cancer. Gynecol Oncol 150(3):545–551. https://doi.org/10.1016/j.ygyno.2018.06.014

Rokavec M, Öner MG, Li H, Jackstadt R, Jiang L, Lodygin D, Kaller M, Horst D, Ziegler PK, Schwitalla S, Slotta-Huspenina J, Bader FG, Greten FR, Hermeking H (2014) IL-6r/STAT3/miR-34a feedback loop promotes EMT-mediated colorectal cancer invasion and metastasis. J Clin Invest 124(4):1853–1867. https://doi.org/10.1172/JCI73531

Sarver AL, Sarver AE, Yuan C, Subramanian S (2018) OMCD: OncomiR cancer database. BMC Cancer 18(1):1223. https://doi.org/10.1186/s12885-018-5085-z

Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD (2003) Asymmetry in the assembly of the RNAi enzyme complex. Cell 115(2):199–208. https://doi.org/10.1016/s00928674(03)00759-1

Segal M, Slack FJ (2020) Challenges identifying efficacious miRNA therapeutics for cancer. Expert Opin Drug Discov 15(9):987–992. https://doi.org/10.1080/17460441.2020.1765770

Sevignani C, Calin GA, Siracusa LD, Croce CM (2006) Mammalian microRNAs: a small world for fine-tuning gene expression. Mamm Genome 17(3):189–202. https://doi.org/10.1007/s00335-005-0066-3

Shah V, Shah J (2020) Recent trends in targeting miRNAs for cancer therapy. J Pharm Pharmacol 72(12):1732–1749. https://doi.org/10.1111/jphp.13351

Shah MY, Ferrajoli A, Sood AK, Lopez-Berestein G, Calin GA (2016) microRNA therapeutics in cancer – an emerging concept. EBioMedicine 12:34–42. https://doi.org/10.1016/j.ebiom.2016.09.017

Shu J, Silva BVRE, Gao T, Xu Z, Cui J (2017) Dynamic and modularized MicroRNA regulation and its implication in human cancers. Sci Rep 7(1):13356. https://doi.org/10.1038/s41598-017-13470-5

Shuang T, Shi C, Chang S, Wang M, Bai CH (2013) Downregulation of miR-17~92 expression increase paclitaxel sensitivity in human ovarian carcinoma SKOV3-TR30 cells via BIM instead of PTEN. Int J Mol Sci 14(2):3802–3816. https://doi.org/10.3390/ijms14023802

Shukuya T, Ghai V, Amann JM, Okimoto T, Shilo K, Kim TK, Wang K, Carbone DP (2020) Circulating MicroRNAs and extracellular vesicle-containing MicroRNAs as response biomarkers of anti-programmed cell death protein 1 or programmed death-ligand 1 therapy in NSCLC. J Thorac Oncol 15(11):1773–1781. https://doi.org/10.1016/j.jtho.2020.05.022

Siddika T, Heinemann IU (2021) Bringing MicroRNAs to light: methods for MicroRNA quantification and visualization in live cells. Front Bioeng Biotechnol 8. https://doi.org/10.3389/fbioe.2020.619583, URL https://www.frontiersin.org/articles/10.3389/fbioe.2020.619583/full

Skog J, Würdinger T, van Rijn S, Meijer DH, Gainche L, Sena-Esteves M, Curry WT, Carter BS, Krichevsky AM, Breakefield XO (2008) Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. Nat Cell Biol 10(12):1470–1476. https://doi.org/10.1038/ncb1800

Smalheiser NR (2007) Exosomal transfer of proteins and RNAs at synapses in the nervous system. Biol Direct 2:35. https://doi.org/10.1186/1745-6150-2-35

Sohel MMH (2020) Circulating microRNAs as biomarkers in cancer diagnosis. Life Sci 248:117473. https://doi.org/10.1016/j.lfs.2020.117473

Song H, Zhang Y, Liu N, Zhang D, Wan C, Zhao S, Kong Y, Yuan L (2016) Let-7b inhibits the malignant behavior of glioma cells and glioma stem-like cells via downregulation of e2f2. J Physiol Biochem 72(4):733–744. https://doi.org/10.1007/s13105016-0512-6

Souza KCB, Evangelista AF, Leal LF, Souza CP, Vieira RA, Causin RL, Neuber AC, Pessoa DP, Passos GAS, Reis RMV, Marques MMC (2019) Identification of cellfree circulating MicroRNAs for the detection of early breast cancer and molecular subtyping. J Oncol 2019:8393769. https://doi.org/10.1155/2019/8393769

Spengler RM, Oakley CK, Davidson BL (2014) Functional microRNAs and target sites are created by lineage-specific transposition. Hum Mol Genet 23(7):1783–1793. https://doi.org/10.1093/hmg/ddt569

Stäehler CF, Keller A, Leidinger P, Backes C, Chandran A, Wischhusen J, Meder B, Meese E (2012) Whole miRNome-wide differential co-expression of microRNAs. Genomics Proteomics Bioinformatics 10(5):285–294. https://doi.org/10.1016/j.gpb.2012.08.003

Sun X, Xu C, Tang SC, Wang J, Wang H, Wang P, Du N, Qin S, Li G, Xu S, Tao Z, Liu D, Ren H (2016) Let-7c blocks estrogen-activated wnt signaling in induction of self-renewal of breast cancer stem cells. Cancer Gene Ther 23(4):83–89. https://doi.org/10.1038/cgt.2016.3

Tan W, Liu B, Qu S, Liang G, Luo W, Gong C (2018) MicroRNAs and cancer: key paradigms in molecular therapy. Oncol Lett 15(3):2735–2742. https://doi.org/10.3892/ol.2017.7638

Théry C, Zitvogel L, Amigorena S (2002) Exosomes: composition, biogenesis and function. Nat Rev Immunol 2(8):569–579. https://doi.org/10.1038/nri855

Tokar T, Pastrello C, Rossos AEM, Abovsky M, Hauschild AC, Tsay M, Lu R, Jurisica I (2018) mirDIP 4.1—integrative database of human microRNA target predictions. Nucleic Acids Res 46:D360–D370. https://doi.org/10.1093/nar/gkx1144

Treiber T, Treiber N, Meister G (2019) Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. Nat Rev Mol Cell Biol 20(1):5–20. https://doi.org/10.1038/s41580-018-0059-1

Tu Y, Liu L, Zhao D, Liu Y, Ma X, Fan Y, Wan L, Huang T, Cheng Z, Shen B (2015) Overexpression of miRNA-497 inhibits tumor angiogenesis by targeting VEGFR2. Sci Rep 5:13827.https://doi.org/10.1038/srep13827

Tüfekci KU, Oner MG, Meuwissen RLJ, Genç S (2014) The role of microRNAs in human diseases. Methods Mol Biol 1107:33–50. https://doi.org/10.1007/978-1-62703-7488_3

Tutar L, Tutar E, Tutar Y (2014) MicroRNAs and cancer; an overview. Curr Pharm Biotechnol 15(5):430–437. https://doi.org/10.2174/1389201015666140519095304

Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO (2007) Exosomemediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9(6):654–659. https://doi.org/10.1038/ncb1596

Valihrach L, Androvic P, Kubista M (2020) Circulating miRNA analysis for cancer diagnostics and therapy. Mol Aspects Med 72:100825. https://doi.org/10.1016/j.mam.2019.10.002

van Dongen S, Abreu-Goodger C, Enright AJ (2008) Detecting microRNA binding and siRNA off-target effects from expression data. Nat Methods 5(12):1023–1025. https://doi.org/10.1038/nmeth.1267

Vinther J, Hedegaard MM, Gardner PP, Andersen JS, Arctander P (2006) Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. Nucleic Acids Res 34(16):e107. https://doi.org/10.1093/nar/gkl590

Wang J, Chen J, Chang P, LeBlanc A, Li D, Abbruzzesse JL, Frazier ML, Killary AM, Sen S (2009) MicroRNAs in plasma of pancreatic ductal adenocarcinoma patients as novel blood-based biomarkers of disease. Cancer Prev Res (Phila) 2(9):807–813. https://doi.org/10.1158/1940-6207.CAPR-09-0094

Wang D, Qiu C, Zhang H, Wang J, Cui Q, Yin Y (2010) Human microRNA oncogenes and tumor suppressors show significantly different biological patterns: from functions to targets. PLoS One 5(9). https://doi.org/10.1371/journal.pone.0013067

Wang H, Peng R, Wang J, Qin Z, Xue L (2018) Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. Clin Epigenetics 10:59. https://doi.org/10.1186/s13148-018-0492-1

Witkos T, Koscianska E, Krzyzosiak W (2011) Practical aspects of microRNA target prediction. Curr Mol Med 11(2):93–109. https://doi.org/10.2174/156652411794859250, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3182075/

Wong N, Wang X (2015) miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res 43:D146–D152. https://doi.org/10.1093/nar/gku1104

Wu G, Dawson E, Duong A, Haw R, Stein L (2014) ReactomeFIViz: a cytoscape app for pathway and network-based data analysis. F1000Res 6:146. https://doi.org/10.12688/f1000research.4431.2, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4184317/

Yang L, Li N, Wang H, Jia X, Wang X, Luo J (2012) Altered microRNA expression in cisplatin-resistant ovarian cancer cells and upregulation of miR-130a associated with MDR1/p--glycoprotein-mediated drug resistance. Oncol Rep 28(2):592–600. https://doi.org/10.3892/or.2012.1823

Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 17(24):3011–3016. https://doi.org/10.1101/gad.1158803

Yu B, Zhao X, Lee LJ, Lee RJ (2009) Targeted delivery systems for oligonucleotide therapeutics. AAPS J 11(1):195–203. https://doi.org/10.1208/s12248-009-9096-1

Yu G, Wang LG, Han Y, He QY (2012) clusterProfiler: an r package for comparing biological themes among gene clusters. OMICS 16(5):284–287. https://doi.org/10.1089/omi.2011.0118, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339379/

Zeka F, Decock A, Van Goethem A, Vanderheyden K, Demuynck F, Lammens T, Helsmoortel HH, Vermeulen J, Noguera R, Berbegall AP, Combaret V, Schleiermacher G, Laureys G, Schramm A, Schulte JH, Rahmann S, Bienertová-Vašků J, Mazánek P, Jeison M, Ash S, Hogarty MD, Moreno-Smith M, Barbieri E, Shohet J, Berthold F, Van Maerken T, Speleman F, Fischer M, De Preter K, Mestdagh P, Vandesompele J (2018) Circulating microRNA biomarkers for metastatic disease in neuroblastoma patients. JCI Insight 3(23). https://doi.org/10.1172/jci.insight.97021

Zhang H, Qi M, Li S, Qi T, Mei H, Huang K, Zheng L, Tong Q (2012) microRNA-9 targets matrix metalloproteinase 14 to inhibit invasion, metastasis, and angiogenesis of neuroblastoma cells. Mol Cancer Ther 11(7):1454–1466. https://doi.org/10.1158/15357163.MCT-12-0001

Zhou H, Rigoutsos I (2014) MiR-103a-3p targets the 5′ UTR of GPRC5a in pancreatic cells. RNA 20(9):1431–1439. https://doi.org/10.1261/rna.045757.114

Zhou K, Liu M, Cao Y (2017) New insight into microRNA functions in cancer: oncogene-microRNA-tumor suppressor gene network. Front Mol Biosci 4:46. https://doi.org/10.3389/fmolb.2017.00046

Zhu X, Li Y, Shen H, Li H, Long L, Hui L, Xu W (2013) miR-137 restoration sensitizes multidrug-resistant MCF-7/ADM cells to anticancer agents by targeting YB-1. Acta Biochim Biophys Sin (Shanghai) 45(2):80–86. https://doi.org/10.1093/abbs/gms099

Zhu J, Xu Y, Liu S, Qiao L, Sun J, Zhao Q (2020) MicroRNAs associated with colon cancer: new potential prognostic markers and targets for therapy. Front Bioeng Biotechnol 8:176. https://doi.org/10.3389/fbioe.2020.00176

Chapter 15
Oxidative Stress, DNA Damage, and Transcriptional Expression of DNA Repair and Stress Response Genes in Diabetes Mellitus

Jéssica Ellen B. F. Lima, Natália C. S. Moreira, Paula Takahashi, Danilo J. Xavier, and Elza T. Sakamoto-Hojo

15.1 Introduction

Diabetes Mellitus (DM) is a known important worldwide public health problem with the number of cases increasing every year. According to the International Diabetes Federation, there were approximately 463 million people between the ages of 20 and 79 years with diabetes worldwide in 2019. This number is projected to reach 578 million by 2030, and 700 million by 2045 (International Diabetes Federation 2019). The two major forms of DM are type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM). T1DM is an autoimmune disorder that occurs predominantly in young individuals, resulting from a destruction of the insulin-producing β-cells by immune cells, which culminates in insulin deficit. T2DM is more common in middle-aged people or above (Newsholme et al. 2016).

Therefore, DM is a chronic metabolic disease that arises from a deficiency of the organism in insulin secretion and/or insulin resistance in the peripheral tissues, which, in turn, leads to chronic high blood glucose levels or hyperglycemia. Ultimately, chronic hyperglycemia has been implicated in long-term complications involving a variety of organs, including kidneys, eyes, heart, nerves, and blood vessels (International Diabetes Federation 2019). Individuals are diagnosed with

J. E. B. F. Lima · N. C. S. Moreira · P. Takahashi · D. J. Xavier
Department of Genetics, Ribeirão Preto Medical School, University of São Paulo,
Ribeirão Preto, São Paulo, Brazil

E. T. Sakamoto-Hojo (✉)
Department of Genetics, Ribeirão Preto Medical School, University of São Paulo,
Ribeirão Preto, São Paulo, Brazil

Department of Biology, Faculty of Philosophy, Sciences and Letters of Ribeirão Preto,
University of São Paulo, Ribeirão Preto, São Paulo, Brazil
e-mail: etshojo@usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_15

diabetes when displaying the following clinical characteristics: glycated hemoglobin levels (HbA1C) ≥6.5% (48 mmol/mol), fasting plasma glucose levels (FPG) ≥126 mg/dL (7.0 mmol/L), 2-h plasma glucose levels after 75 g glucose load ≥200 mg/dL (11.1 mmol/L) (in the absence of unequivocal hyperglycemia, these three parameters should be confirmed by retaking the test), or for individuals with classic hyperglycemic symptoms/hyperglycemic crisis, casual plasma glucose levels ≥200 mg/dL (11.1 mmol/L) (International Diabetes Federation 2019).

There is evidence of an association between oxidative stress and both types of DM. Interestingly, while oxidative stress can be a consequence of these disorders due to hyperglycemia, it can also be a contributing factor to the pathogenesis of both T1DM and T2DM, considering that reactive molecules play a crucial role in pancreatic β-cell damage, in addition to the damage to other tissues and biological macromolecules.

15.1.1 Oxidative Stress

The advancement of age imposes a progressive increase in the production of reactive oxygen species (ROS), and also a reduced antioxidant defense capacity of the body (Cui et al. 2012). ROS encompass a group of reactive molecules produced from molecular oxygen by reduction–oxidation (redox) reactions. These molecules can be divided into non-radicals and free radical species. The superoxide anion radical ($O_2^{\bullet-}$) and the hydroxyl radical ($OH^{\bullet}$) are commonly referred to as free radicals because they have at least one free electron. Hydrogen peroxide (H_2O_2), singlet molecular oxygen ($^1O_2$), and hypochlorous acid (HOCl) are some major known non-radical species (Collin 2019; Sies and Jones 2020). ROS are normal byproducts of cellular metabolism generated by various endogenous and exogenous sources. The main endogenous sources are the mitochondrial electron transport chain and nicotinamide adenine dinucleotide phosphate (NADPH) oxidases. However, ROS can also be produced in different cell compartments (such as the endoplasmic reticulum [ER], lysosomes, peroxisomes, and membranes) by the action of various stimuli (nutrients, growth factors, hormones, cytokines, and others) (He et al. 2017; Sies and Jones 2020). In addition to intracellular endogenous sources, ROS can be generated by exogenous physical agents (ultraviolet light, X-rays, and γ-rays), air pollutants, tobacco smoke, heavy metals, and certain drugs (Moldogazieva et al. 2019).

Under normal conditions and appropriate levels, ROS play important roles in several physiological processes, participating in cell signaling pathways, glucose uptake, memory function, and cell–cell interactions (Collin 2019; Sies and Jones 2020). On the opposite, although moderate amounts of ROS have positive effects (such as killing of invading pathogens, wound healing, and repairing processes), the misregulation of ROS production can cause oxidative damage to macromolecules, as well as mitochondrial dysfunction and cell death (Fig. 15.1) (Peoples et al. 2019). When present at very high concentrations, ROS react with lipids, proteins, and DNA, and can be severely detrimental to cells, thus causing a condition of redox imbalance. This condition leads to the concept of oxidative stress that arises from the imbalance between prooxidant species and the antioxidant defense (Moldogazieva et al. 2019; Sies and Jones 2020). Moreover, given the fact that superoxide anion and hydrogen peroxide levels are significantly higher in the mitochondrial matrix than in cytosolic and nuclear spaces, mitochondria become primary targets for ROS-induced damage (Luo et al. 2020). The excessive ROS can react with several mitochondrial proteins, compromising its function and dynamics and also disrupt the mitochondrial permeability transition pore, which facilitates the escape of electrons from the electron transport chain generating ROS and directly contributing to the release of ROS to the cytosol (He et al. 2017; Sies and Jones 2020). Mitochondrial dysfunction affects several biological processes, including nuclear genomic stability and cellular bioenergetics (Luo et al. 2020). Besides, it have been associated with several diseases, such as diabetes (Bhansali et al. 2017) and neurodegenerative diseases (Delbarba et al. 2016).

The constant exposure to oxidants triggers many enzymatic and nonenzymatic mechanisms, as well as adaptive responses to counteract ROS, ultimately leading to the reestablishment of the cellular oxidant/antioxidant homeostasis (Kalyanaraman 2013). The enzymatic mechanisms include the action of superoxide dismutase (SOD), catalase (CAT), glutathione peroxidase (GPx), and thioredoxin system. SOD enzymes (cytosolic copper/zinc SOD, mitochondrial manganese SOD, and extracellular SOD) are involved in the normal dismutation of $O_2^{\bullet-}$ in O_2 and H_2O_2; H_2O_2 can be detoxified in $H_2O + O_2$ mainly by CAT, when H_2O_2 levels are low, and by GPx when these levels are high. GPx also metabolizes other lipid peroxides (LOOH) (Fig. 15.1) (Kalyanaraman 2013; He et al. 2017). Regarding the mechanism of adaptive response to ROS, there is a signaling pathway under cell exposure to the oxidative challenge, with the activation of genes encoding antioxidant enzymes responsible for the maintenance of redox homeostasis in eukaryotes. A crucial gene is the nuclear factor erythroid 2-related factor 2 gene (*NRF2*), which is a master regulator of the antioxidant response (He et al. 2020).

On the other hand, there are also nonenzymatic detoxification mechanisms, which include the small molecular weight antioxidants: ascorbic acid (vitamin C), α-tocopherol (vitamin E), reduced glutathione (GSH), carotenoids, lycopene, some minerals (zinc, manganese, and selenium), omega-3 and 6, flavonoids, among others. While vitamin C reacts rapidly with several ROS types, such as superoxide, hydrogen peroxide, and hydroxyl radical, vitamin E can halt lipid peroxidation (He et al. 2017).

Nevertheless, stress conditions that impair those adaptive mechanisms, reduce the concentrations of antioxidants, or affect the action of antioxidant enzymes can lead to oxidative stress by disrupting the redox homeostasis due to excessive ROS production, leading to ROS-mediated damage of important organelles and biomolecules (He et al. 2017). Therefore, when there is a redox imbalance, in which the levels of prooxidants exceed those of antioxidants, a consequent induction of damage to macromolecules can occur, such as lipid peroxidation, protein oxidation, and DNA damage, leading to disease (Kalyanaraman 2013; He et al. 2017), as illustrated in Fig. 15.1.

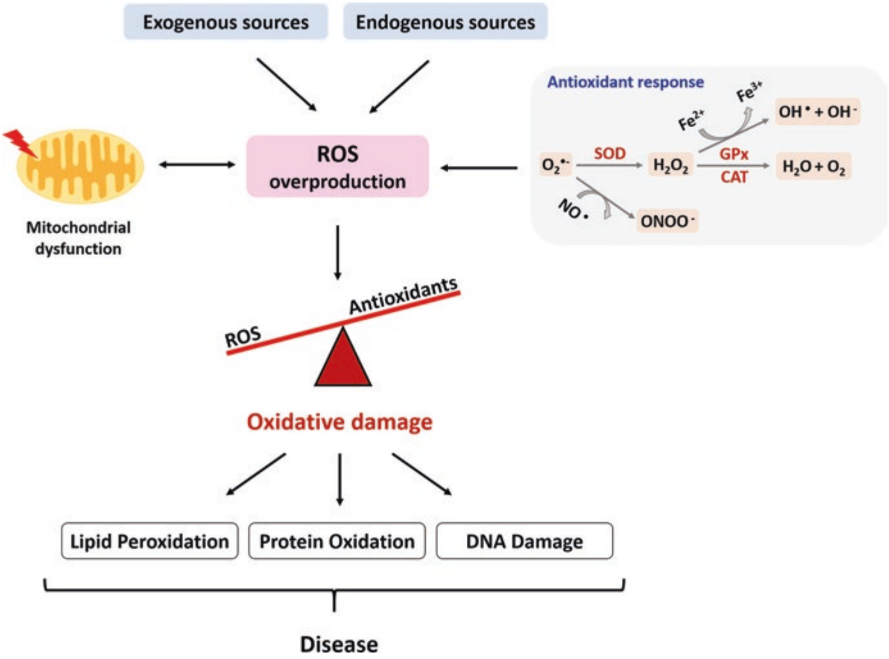


Fig. 15.1 ROS and oxidative damage. ROS can be generated by endogenous and exogenous sources. There are antioxidant mechanisms that maintain ROS homeostasis; however, overproduction of ROS can occur causing a prooxidant redox imbalance, which leads to oxidative damage and mitochondrial dysfunction. Furthermore, mitochondrial dysfunction increases even more ROS production. Excessive ROS levels can induce lipid peroxidation, protein oxidation, and DNA damage, thus leading to disease. ROS: reactive oxygen species; $O^{\bullet-}$: superoxide radical; $OH^{\bullet}$: hydroxyl radical; OH^-: hydroxyl ions; H_2O_2: hydrogen peroxide; O^-: superoxide radical; $ONOO^-$: peroxynitrite; $NO^{\bullet}$: nitric oxide; SOD: superoxide dismutase; GPx: glutathione peroxidase enzyme; CAT: catalase

In the DNA, ROS can react with both purines and pyrimidines, generating a number of modified DNA base products. Guanine exhibits a low redox potential and, it is preferentially oxidized, with 8-oxo-7,8-dihydroguanine (8-oxoguanine; 8-oxoG) being the most extensively oxidized DNA lesion that has been measured in several studies (Storr et al. 2013; Dizdaroglu et al. 2017). Nucleotides are also prone to oxidation by ROS, and the oxidized bases, deoxynucleotide triphosphates, (especially oxidized dGTP and dATP), could be erroneously incorporated into the DNA during the replication (Sun et al. 2015). Following DNA replication, these lesions can cause a change from GC to TA (transversion), thus causing mutations (Włodarczyk and Nowicka 2019). Therefore, DNA damage can exert an impact on DNA replication fidelity, leading to mutations, and imposing a risk to cell metabolism and survival (Włodarczyk and Nowicka 2019). ROS not only induce oxidized bases, abasic (AP) sites, and single-strand breaks (SSBs) (Hegde et al. 2012), but can also cause DNA intrastrand and interstrand crosslinks, DNA–protein crosslinks, double-strand breaks (DSBs), as well as damage to the DNA sugar moiety (Storr et al. 2013; Dizdaroglu et al. 2017).

DNA repair mechanisms play a crucial role in the maintenance of genome integrity against a large variety of endogenous and exogenous ROS, and also chemical and physical agents, guaranteeing the fidelity of DNA replication, thus preventing mutations and their consequences (Hanawalt and Wilson 2016). Base excision repair (BER) is a well-known major mechanism involved in the repair of ROS-induced oxidative lesions and SSBs in the DNA. BER requires several proteins; the key enzymes are DNA glycosylases, which remove different damaged bases by cleavage of the N-glycosylic bonds between the bases and the deoxyribose moieties of the nucleotide residues. Briefly, several enzymes participate in a sequential pathway, performing the initial lesion recognition, removal of oxidized bases by DNA glycosylases, such as 8-oxoguanine DNA glycosylase (OGG1), followed by the DNA incision by the apurinic/apyrimidinic endonuclease 1 (APE1), and subsequent recruitment of DNA polymerase β (Pol β), which perform gap filling; subsequently, ligase 1 or a complex of X-ray repair complementing protein 1 (XRCC1) and ligase IIIα seals the resulting gap to complete the repair process (Svilar et al. 2011; Krokan and Bjørås 2013; Cadet and Davies 2017).

BER is the major pathway for the repair of oxidative base damage, but other repair processes exist in eukaryotes, such as nucleotide excision repair (NER), mismatch DNA repair (MMR), translesion synthesis (TS), homologous recombination (HR), and nonhomologous end-joining (NHEJ); these repair processes are important mechanisms implicated in the repair of several types of DNA lesions (base damages, SSBs, DSBs, crosslinks, adducts, intercalation, among others) (Krokan and Bjørås 2013; Cadet and Davies 2017), being integrated with several cellular processes, such as cell cycle regulation, apoptosis, transcription, and replication. They may also be activated in response to oxidative damage, as an alternative repair pathway (Slupphaug 2003; Surova and Zhivotovsky 2013). Interestingly, Souza-Pinto et al. (2009) demonstrated that RAD52 (recombination protein) cooperates with OGG1 to repair oxidative DNA damage, suggesting a coordinated action between these proteins.

15.1.2 Diabetes Mellitus and Oxidative Stress

Hyperglycemia may lead to increased oxidative stress by the direct production of ROS, or by changes in the redox homeostasis through the disruption of a variety of mechanisms. The production of ROS in DM can be induced by endothelial and vascular smooth muscle cells, NADPH oxidase, xanthine oxidase, cyclooxygenase, and uncoupled NOS, while nonenzymatic sources include the generation of superoxide by the mitochondrial respiratory chain, advanced glycation end products (AGEs), activation of protein kinase C, glucose autoxidation process, and activated polyol pathway (Ahmad et al. 2017). All of these pathways can trigger redox imbalance (elevated oxidative stress), leading to damaged macromolecules that

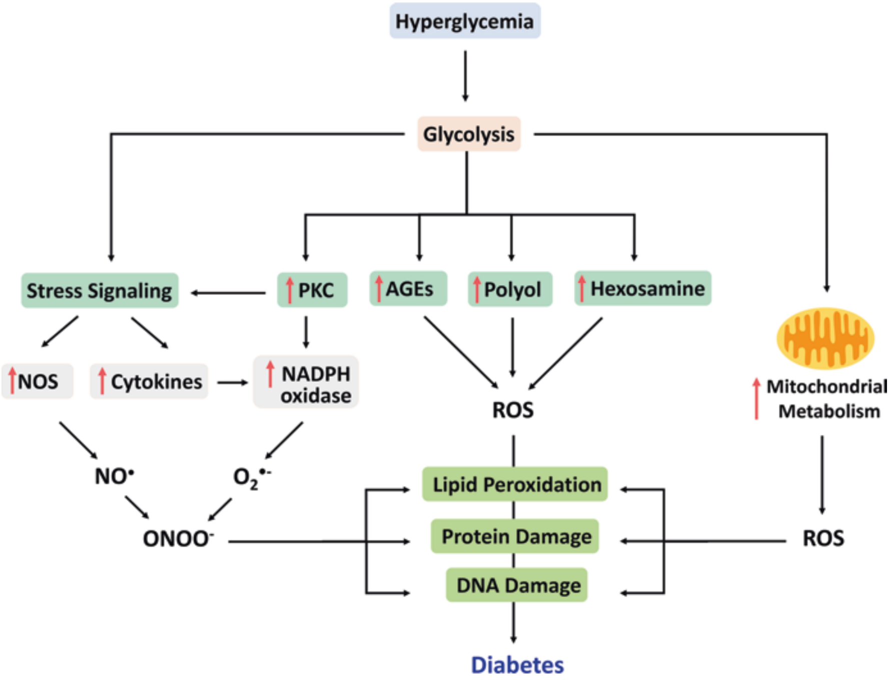


Fig. 15.2 Hyperglycemia-induced oxidative stress. Oxidative stress is a critical contributing factor to diabetes. A chronic state of hyperglycemia can promote an increase in glycolysis, which, in turn, acts on several signaling pathways, such as polyol, advanced glycosylation end products (AGEs), hexosamine, and protein kinase C (PKC) pathways. All of these pathways can produce oxidative stress. Consequently, the stress signaling and the activation of mitochondrial metabolism lead to increased nitric oxide synthases (NOS), cytokines, and NADPH oxidase and generation of superoxide by the mitochondrial respiratory chain. These pathways, when deregulated, promote the production of ROS, redox imbalance (a condition of oxidative stress), increasing lipid peroxidation, damage to proteins and DNA, consequently leading to diabetes. $O^{\bullet -}$: superoxide radical; ONOO−: peroxynitrite; $NO^{\bullet}$: nitric oxide

ultimately, are implicated in the development and progression of DM (Fig. 15.2) (Sifuentes-Franco et al. 2017).

T1DM and T2DM are metabolic disorders, apparently with distinct mechanisms, but in both diseases, there is a significant loss of insulin-producing β-cells due to cell death. High acute or chronic glucose levels in diabetic patients promote an increase in the glycolytic flux and, consequently, an increase in the mitochondrial metabolism, which exacerbate ROS production (Gerber and Rutter 2017; Volpe et al. 2018). Thus, hyperglycemia induces excessive production of superoxide anion in the mitochondrial electron transport chain, formation of AGEs, as well secretion of proinflammatory cytokines, leading to numerous metabolic abnormalities and activation of pathways involved in the development of DM and diabetes complications (Djeli et al. 2019). Particularly, ROS are associated with the activation of inflammatory pathways that lead to apoptosis in β-cells (Hurrle and Hsu

2017), reduction of glucose transporters, inhibition of insulin receptor, and decreased insulin gene expression, which are associated with insulin resistance (Boucher et al. 2014; DeFronzo et al. 2015). This picture is compatible with alterations in gene expression profiles, as observed in peripheral blood mononuclear cells (PBMCs) of diabetic patients who display chronic hyperglycemia (Xavier et al. 2015).

15.2 Type 1 Diabetes Mellitus

T1DM is a consequence of the autoimmune destruction of the insulin-producing pancreatic β-cells, which eventually ceases insulin production and hence the glucose uptake by the tissues of the body, and culminates in hyperglycemia (Katsarou et al. 2017). Approximately 5–10% of all diabetic patients have T1DM, which can occur at any age, although it usually arises during childhood and adolescence (International Diabetes Federation 2019). The patients with T1DM commonly present classical symptoms as ketoacidosis, excessive thirst, blurred vision, bedwetting, frequent urination, fatigue, constant hunger, and weight loss (International Diabetes Federation 2019). They require daily insulin injections to control glucose levels and avoid life-threatening hypoglycemia; metformin, glucagon-like peptide-1 receptor (GLP4-1R) agonists, sodium-glucose cotransporter-2 (SGLT2) inhibitors, dipeptidyl peptidase-4 (DPP4) inhibitors may also be used in some cases (International Diabetes Federation 2019).

The exact cause of T1DM has not been elucidated, but it might be a result of a complex combination of several susceptibility genes (with functions on metabolism and immune system), as well as environmental factors (DiMeglio et al. 2018). Over 90% of T1DM cases have the characteristic presence of autoantibodies for glutamate decarboxylase (GAD65), islet antigen-2 (IA-2), tetraspanin-7, and zinc transporter 8 (ZNT8) that have been usually used as biomarkers of the disease (Katsarou et al. 2017; DiMeglio et al. 2018).

Individuals with the HLA class 2 haplotypes, *HLA DRB1*0301-DQA1*0501-DQ*B10201 (DR3)* and *HLA DRB1*0401-DQA1*0301-DQB1*0301 (DR4-DQ8),* located on chromosome 6, have the highest genetic risk for T1DM, which are related to the development of β-cell-targeted autoimmunity (Katsarou et al. 2017); in addition, non-HLA genes including those encoding insulin (*INS*), cytotoxic T-lymphocyte-associated protein 4 (*CTLA4*), non-receptor protein tyrosine phosphatase type 22 (*PTPN22*), and interleukin 2 receptor alpha (*IL2RA*) have been identified by genome-wide association studies (GWAS) as having strong associations with the disorder (Nyaga et al. 2018). Regarding environmental factors, viral infections or toxins, as well as climate and diet have been suggested to contribute to T1DM onset (International Diabetes Federation 2019).

15.2.1 Oxidative Stress and DNA Damage in T1DM Patients

There is evidence that oxidative stress plays a crucial role in the increased inflammation and release of cytokines, which promotes the destruction of β-cells and the development of T1DM (Nassima et al. 2014). In this context, it has been investigated the levels of antioxidants, markers of oxidative stress, and DNA damage in patients suffering from T1DM relative to healthy subjects.

Studies with T1DM children and adults have consistently shown an increased oxidative stress condition, evidenced by increased total antioxidant status (TAS) (Gheni et al. 2020), high levels of lipid peroxidation (malondialdehyde (MDA), lipoperoxides (LPO), and 8-iso-prostaglandin F2α (8-iso-PGF2α)), protein oxidation, DNA oxidation (8-OhDG), and carbonylated proteins (Goodarzi et al. 2010; Codoñer-Franch et al. 2010; Nassima et al. 2014; Altincik et al. 2016). Moreover, it has been reported that glucose fluctuations may potentiate oxidative stress in non-obese T1DM children (Meng et al. 2015). On the other hand, the activity of antioxidant enzymes (GPx, glutathione reductase (GR) and SOD), in addition to lipophilic antioxidants (α-tocopherol and β-carotene), was found significantly decreased in T1DM patients compared to healthy subjects (Codoñer-Franch et al. 2010; Nassima et al. 2014). In fact, at the early onset of the disease, T1DM children already show glutathione depletion compared to healthy subjects (Pastore et al. 2012).

Furthermore, it has been reported an association between increased oxidative stress and impaired antioxidant status with lower levels of Magnesium and Zinc (Zn) and increased levels of Copper (Cu), in particular, in poorly controlled (with HbA1c $\geq 9\%$) children (Salmonowicz et al. 2014) and adults (Lin et al. 2014) with T1DM. Abnormalities in Zn and Cu levels seem to be associated with increased oxidative stress and diabetes complications (Bjørklund et al. 2019). Interestingly, Salmonowicz et al. (2014) and Lin et al. (2014) found an increase in CAT and SOD activity in T1DM patients, respectively, which might be compensating for the excessive ROS generation in these individuals, but the lower total antioxidant status might indicate a deficiency of the antioxidant system in those patients. Moreover, the disruption of thiol/disulfide homeostasis, known to play an important role in antioxidant responses, is also associated with oxidative damage, and it is considered another approach to evaluate oxidative stress (Ates et al. 2016). Durmus et al. (2019) and Ates et al. (2016) have shown that T1DM patients have a shift of thiol/disulfide homeostasis toward disulfide direction, an indication of oxidative stress, associated with chronic inflammation, which is followed by high levels of c-reactive protein, besides hyperglycemia.

Additionally, it has been demonstrated that men and women with T1DM have significantly more DNA damage and oxidized DNA damage (measured by Fpg-sensitive sites) in comparison with their corresponding controls (Dinçer et al. 2003). Even those T1DM patients with acceptable glycemic control reported significantly elevated rates of DNA damage (Hannon-Fletcher et al. 2000). Chronic hyperglycemia-induced cellular damage and oxidative stress are strongly associated with micro- and macrovascular complications of diabetes. Recent studies have shown a correlation between increased oxidative stress (high MDA and nitric oxid (NO) levels) and hematologic alterations in T1DM patients (Abdel-Moneim et al. 2020). Abnormalities in erythrocytes have been suggested to play a pivotal role in the development of microvascular complications (Abdel-Moneim et al. 2020). High levels of lipid peroxidation and NO have been linked to greater severity of retinopathy (Ruia et al. 2016). The glycemic control and oxidative stress management is an important issue to be addressed. It has been reported that glycemic variability in T1DM patients induces alterations in erythrocyte membrane stability (Rodrigues et al. 2018); recently, a correlation has been found between low levels of antioxidants enzymes (SOD and glutathione), high levels of oxidative stress (MDA), and impairment of bone formation, in children with T1DM (El Amrousy et al. 2021). Collectively, these studies suggest an impairment of the antioxidant defense system and an increase in oxidative stress and DNA damage in T1DM patients, which are associated with T1DM progression and later complications.

15.2.2 Transcriptional Expression Profiles of Oxidative Stress and DNA Repair Genes in T1DM Patients

A wide interest has been directed to the investigation of molecular pathways underlying the development and progression of T1DM and T2DM and their interactions. Within this context, studies at a genomic scale have been providing a large volume of data that can help in understanding and clarifying the etiopathogenesis of diabetes, with the possibility of contributing to the development of new therapeutic strategies.

Since the initial work reported by Schena et al. (1995), the microarray technique became a common and important tool in medical and biological research. Over more than two decades, several studies developed at a large scale regarding transcriptional profiling have been performed to compare expression profiles displayed by patients (for several diseases) relative to healthy subjects. Recently, Gastol et al. (2020) have performed a microarray study in T1DM patients to look for alterations in molecular pathways, by analyzing differentially expressed genes (DEGs) in blood cells. Interestingly, they found that T1DM patients showed an upregulation of genes related to DNA repair (*APEX1, ERCC3, ERCC5, PARP1, PARP4, MLH1, XPC*), antioxidant enzymes (*PRDX1, SOD1, SOD2*), ER-stress response (*ATF6, PRDX6, GCLC, TXNRD1*), proteasome and autophagosome formation (*ATG3, ULK1, BECN1, DNAJB1, SQSTM1*), apoptosis (caspases, TNF family factors, and their receptors), inflammation (*NFKB, Il-10, Il-1b*), and activation of inflammatory pathways. On the other hand, the patients presented a downregulation of genes involved in glucose transport (*SLC2A11*), glutathione synthesis (*GCLM*), expression of mitochondrial proteins of complexes I and III, and proteolytic enzymes (*cathepsins, FRAP1, ATG10, GABARAPL2*). Possibly, the inhibition of mitochondrial proteins and proteins involved in glutathione synthesis might be responsible for the induction of oxidative stress, which can be related to the activation of DNA repair and antioxidant pathways.

There is evidence that ER stress is involved in the destruction of pancreatic β-cells, triggering the development of both T1DM and T2DM. The ER is a major organelle responsible for regulating protein synthesis, folding, maturation, and transport and has a key role in insulin synthesis. The ER maintains a controlled balance between the synthesis and proper protein folding. However, several conditions can break this homeostatic balance, such as excess of nutrients, insulin resistance, increased levels of ROS, and inflammation related to obesity. The disturbance of this homeostasis leads to an accumulation of misfolded proteins in the organelle, either by an increased rate of protein synthesis or by alterations in the ER *milieu*, compromising the efficiency of protein folding. Regardless of the case, the unfolded protein response (UPR) is triggered to restore protein homeostasis. In patients with diabetes, hyperglycemia triggers the ER to produce an excessive amount of insulin, which overloads the ER, leading to the accumulation of misfolded and unfolded proteins. The ER overload induces a stress condition that activates the UPR, and under pathological conditions and excessive ER stress, the promotion of cell death may occur (Cao et al. 2020). Mainly three proteins are responsible for the activation of UPR: inositol-requiring protein-1α (IRE1α), protein kinase RNA (PKR)-like ER kinase (PERK), and activating transcription factor 6 (ATF6) (Cao et al. 2020). Accordingly, Gastol et al. (2020) observed an increased expression of ATF6, which has been linked to the activation of autophagy (Walter et al. 2018). In addition, it has been suggested that ER stress induces the release of proinflammatory cytokines, which was confirmed by the increased plasma levels of IL-6 and activation of inflammatory genes. The authors also discussed that despite the upregulation of proteasome and autophagosome formation in T1DM, the removal of damaged proteins can be compromised by a concomitant downregulation of lysosomal proteolytic enzymes, which has been associated with ER stress (Cao et al. 2020).

Irvine et al. (2012) investigated whether there were differences in gene expression of purified peripheral blood CD14+ monocytes between recently diagnosed T1DM children and adult healthy controls by applying the whole-genome microarrays, followed by validation of some genes by quantitative polymerase chain reaction (qPCR). The authors showed that the monocyte expression profiles exhibited by the patients were clustered into two subgroups, with one of them (group B) clustering separate from the other patient subgroup and the healthy controls. At diagnosis, both subgroups of patients were clinically identical, however, group B presented increased levels of HbA1c 3 and 6 months after diagnosis and needed significantly higher insulin doses during the first year of the disease. Expression profiles in monocytes from patients belonging to group B showed an upregulation of genes related to the UPR, which results from ER stress (*IRE1*, *GRP78*, *DDIT3*, *XBP1*), *HIF1A*, which is a major mediator of oxidative stress, and several of its targets (*DDIT4*, *PFKFB3*, and *ADM*); while genes that play a role in mitochondrial oxidative phosphorylation (*PDHB*, *MDH1*, *IDH1*, *SDHC*, *ACLY*) and cellular antioxidant pathways (*CAT*, *G6PD*, *OXR1*, *PRDX1*, *PRDX3*) were found downregulated, indicating perturbation of protective systems (Irvine et al. 2012). Moreover, mitochondrion was the most significantly enriched cellular component term for the downregulated genes in group B. The two biological processes, oxidative and ER stresses were found closely associated. Oxidative stress can promote ER stress, and in response to that, ER activates the UPR transcriptional program. UPR failure may lead to prolonged ER stress, which, in turn, triggers apoptosis and inflammation (Cao et al. 2020). Accordingly, genes controlling apoptosis were enriched in monocytes from group B patients. Hence, collectively, these findings imply that the group B monocytes are intrinsically susceptible to stress or exist in a stressful environment, as well as indicate the persistence of ER stress (Irvine et al. 2012).

Intriguingly, Stechova and co-workers (2012) compared gene expression profiles of freshly isolated PMBCs from T1DM patients, their first-degree relatives with higher genetic risk of developing the disease, and nondiabetic individuals by the microarray technology. They observed a clear difference between the expression profiles of relatives of patients (in particular the autoantibody-negative ones) and healthy controls. Moreover, the highest number of differentially activated cell signaling processes (99 pathways), including DNA damage and oxidative stress pathways was reported in the comparison between the relatives, regardless of autoantibody status, and the control group. Thus, these findings showed that nondiabetic relatives of T1DM patients also present alterations in gene expression. Caramori et al. (2015) performed a transcriptional profiling study in skin fibroblasts taken from 100 T1DM patients and found that longstanding T1DM patients (without diabetic nephropathy) displayed upregulation of DNA repair pathways, DNA replication, cell cycle, and RNA degradation compared to T1DM patients with nephropathy, and also compared to healthy controls. The authors suggest that the increased expression of repair pathways may be involved in preventing or delaying the onset of nephropathy.

Another study investigated gene expression profiles of endothelial progenitor cells (EPC), which were *in vitro* differentiated from PBMCs, from T1DM patients pre- and post-supplementation with folic acid (FA, a B-vitamin with antioxidant properties) and nondiabetic individuals (van Oostrom et al. 2009). The authors found 1591 DEGs between pre-FA treatment T1DM patients and the control group. These genes were associated with several processes including response to stress and response to hypoxia. Among the upregulated genes (related to these two terms) detected in EPC from T1DM patients were dual oxidase 2 (*DUOX2*), a NADPH oxidase that can produce superoxide, nitric oxide synthase 2A (*NOS2A*) that is capable of generating NO , thioredoxin reductase 2 (*XNRD2*), a major enzyme involved in the control of the intracellular redox balance, lactoperoxidase (*LPO*) and NADPH oxidase organizer 1 (*NOXO1*), which is associated with the generation of ROS. Importantly, after FA treatment the gene expression profiles (513 of the 1591 DEGs) in diabetic EPC normalized to levels similar to those exhibited by healthy individuals. As expected, FA altered the expression of oxidative stress-associated genes in EPC, with four (*DUOX2*, *NOS2A*, *NOXO1,* and *LPO*) being included among the 513 normalized genes. In addition, another differentially expressed gene (down-regulated) in T1DM patients that was normalized by FA treatment was the transcription factor V-maf musculoaponeurotic fibrosarcoma oncogene homolog F (*MAFF*). This transcription factor can bind to NRF2, which, in turn, plays a crucial role in the antioxidant defense (Golpour et al. 2020).

Recent advances have brought novel high-throughput multi-omics approaches to provide a deep comprehensive understanding between disease state and the molecular profiles of healthy individuals. Instead of a single omics analysis, Balzano-Nogueira et al. (2021) applied an integrative approach to evaluate gene expression profiles, metabolomics, and dietary biomarkers to establish a multi-omics signature in children up to 12 months before T1DM development. Interestingly, before T1DM development, the children displayed upregulation of several genes associated with glucose utilization, energy metabolism, DNA repair, ROS scavenging, ER-protein processing, and apoptosis compared to the children who did not develop T1DM. Additionally, arachidonate-lipoxygenase genes (*ALOX12*, *ALOX15*, *ALOX15B,* and *PTGS1*), which are known to be activated by ROS and to increase the release of proinflammatory and pro-angiogenic molecules, were found upregulated. On the other hand, immune system pathways (regulation of natural killer immunity, CXCR4 signaling, TGFβ signaling, FOXO signaling) became downregulated just before the development of T1DM. At 0–3 months before T1DM diagnosis, pathways associated with antigen presentation (NF-kB signaling and insulin signaling) became strongly activated. Altogether, the authors pointed out molecular profiles following the progression of T1DM, supporting the hypothesis that the increased oxidative stress and inflammation do occur even several months before the onset of T1DM, being related to the increased activity of proinflammatory cytokines, activation of pathways that favors autoimmunity and cellular damage, concomitantly with abnormalities in lipid metabolism and nutrient uptake.

Regarding the expression of noncoding protein genes, microRNAs (miRNAs) have been indicated both as potential biomarkers for the earlier diagnosis of diabetes and as therapeutic targets for the treatment of this disorder (Assmann et al. 2017). MiRNAs are endogenous noncoding RNA molecules of approximately 22 nucleotides that are involved in the posttranscriptional regulation of protein-coding gene expression by base-pairing to specific sites generally in 3’ untranslated regions (UTRs) of the messenger RNA (mRNA) targets; in this way, miRNAs lead to the degradation and/or translational downregulation of their targets (Agbu and Carthew 2021). Takahashi et al. (2014) compared the miRNA expression profiles displayed by PBMCs from T1DM patients with those from healthy nondiabetic controls by performing microarray experiments. The authors identified a set of 44 differentially expressed miRNAs (35 upregulated and nine downregulated) that clearly distinguish T1DM patients from healthy subjects. After target prediction, results pointed to 10,827 and 6,636 potential targets of the up- and downregulated miRNAs, respectively; of note, a total of 85 and 75 genes implicated in DNA repair and response to oxidative stress, respectively, are potential targets of the 44 differentially modulated miRNAs in T1DM. Furthermore, Assmann et al. (2017) performed a systematic review of several miRNA studies performed in different tissues (serum, plasma, PBMCs, or pancreas) from T1DM patients compared to the controls. They found several circulating miRNAs (miR-21-5p, miR-24-3p, miR-148a-3p, miR-181a-5p, miR-210-5p, and miR-375) that were upregulated and some miRNAs (miR-146a-5p, miR-150-5p, miR-342-3p, miR-1275, and miR-100-5p) that were found downregulated in T1DM patients compared to the nondiabetic controls. Regarding the upregulated miR-21-5p, in a previous report in the literature, its function was implicated in anti-inflammatory process by inhibition of the NF-kB signaling pathway (Sheedy et al. 2010); regarding other relevant functions attributed to miRNAs, miR-24 and miR-148 activate insulin expression (Agbu and Carthew 2021), while miR-148a-3p is a regulator of β-cell self-tolerance and autoimmunity (Gonzalez-Martin et al. 2016); and miR-210-5p targets include genes related to mitochondrial metabolism, DNA repair, angiogenesis, and cell survival (Devlin et al. 2011).

Taken together, studies on the whole-transcript expression showed important alterations related to the expression of DNA repair and antioxidant genes, ER stress response, UPR, apoptosis, mitochondrial genes, and inflammation in T1DM patients, as well as in their relatives and also in children before T1DM development. Moreover, those genes are putative targets of a set of miRNAs that clearly distinguished T1DM patients from healthy individuals. These data support the hypothesis that patients with T1DM respond to the increased oxidative stress and DNA lesions by means of changes in their gene expression profiles, which probably, may affect several biological processes, and may explain the physiological alterations in the course of the disease.

15.3 Type 2 Diabetes Mellitus

T2DM is the most common type of DM, accounting for approximately 90% of all diagnosed cases of diabetes (International Diabetes Federation 2019). The disease is mainly characterized by hyperglycemia resulting from resistance to insulin action and/or by a deficiency in the secretion of this hormone, presenting a great correlation with an unhealthy lifestyle, aging, obesity, and lack of physical activity (DeFronzo et al. 2015). The symptoms of T2DM are similar to those of T1DM but less acute or intense. The majority of T2DM cases are symptomless and remained undiagnosed for a long period, until the hyperglycemia starts to trigger a series of complications, including diabetic retinopathy, nephropathy, neuropathy, cardiovascular diseases (International Diabetes Federation 2019), and more recently, it has been reported risk to the development of dementia (Mittal and Katare 2016; Chatterjee and Mudher 2018).

The biochemical mechanisms and physiological processes that characterize T2DM are not well understood. Nevertheless, over 500 genomic regions and some susceptibility genes have been identified by GWAS, including *PPARG*, *KCNJ*, *CAPN10*, *FTO*, *CDKN2A/B*, *CDKAL1*, *TCF7L2,* and *IGFBP2*; in addition, several identified genes are associated with diabetes complications, such as *GJA8* and *SLC18A2* (for retinopathy), *UMOD and TENM3* (for nephropathy), *NRP2* (for neuropathy), and *SORT1* (for coronary heart disease); furthermore, *SCN3A* and *SV2A* genes, which are potential targets for therapeutic purpose, were also identified (Vujkovic et al. 2020). Additionally, some studies have been exploring epigenetic alterations in diabetes, since they may also contribute to the genetic susceptibility to T2DM (Basile et al. 2014; Kwak and Park 2016).

The main metabolic alterations in T2DM are insulin resistance, β-cell dysfunction, and chronic inflammation (DeFronzo et al. 2015). Obesity is a great contributor to those alterations, causing a chronic inflammatory response in the adipose tissue, characterized by abnormal production of cytokines, which include mostly molecules playing roles in stress response processes and activation of inflammatory pathways, such as the JNK and NF-kB pathways (Tsalamandris et al. 2019). Among the released cytokines are the Tumor Necrosis Factor-alpha (TNF-α) and Interleukin 6 (IL-6), which are released at large amounts by adipocytes and act inhibiting the tyrosine phosphorylation of insulin receptor substrate (IRS-1) impairing the insulin signaling pathway and leading to insulin resistance (Chen et al. 2017). Besides, excessive body fat leads to increased circulation of free fatty acids (FFA), known to impair pancreatic β-cell function and decrease insulin secretion, in addition to increasing the release of TNF-α and IL-6. Under this condition, there is a preferential use of lipids as an energy source, especially by muscles, which prevents glucose utilization and glycogen synthesis, leading to hyperglycemia. Furthermore, there is an increase in insulin secretion to compensate for the insulin receptor resistance, and this condition gradually leads to the development of the disease (Huang et al. 2018). Interestingly, a whole-blood transcriptome study in a large cohort (comprising 1977 nondiabetic obese subjects) reported a correlation between increased body mass index (BMI) and downregulation of several genes involved in insulin signaling (*IRS2*, *PIK3CD*, *PIK3R4*, *PDPK1*, *AKT1, PTEN, PTPN1*), DNA repair (*ATM*) and defense against ROS, including target genes and regulators of NRF2 (SOD2, *NFE2L2, TXNRD1, MGST2, GSTM2, NQO2*), suggesting that these alterations may contribute to T2DM development in obese subjects (Homuth et al. 2015).

As already mentioned, in diabetes, the chronic increase in glucose levels leads to overproduction of ROS and AGEs, and consequently, oxidative stress. ROS and oxidative stress activate pathways linked to increased release of proinflammatory cytokines, growth factors, adhesion molecules, and procoagulant factors, all culminating in β-cell dysfunction, insulin resistance, endothelial dysfunction, and T2DM progression with micro- and macrovascular complications (Akash et al. 2013).

Currently, the first line of therapy for T2DM patients is the recommendation of changes in lifestyle concomitantly with diet and weight management, in addition to regular physical activity and the use of glucose-lowering medicaments, such as metformin, which is the preferred drug as the initial pharmacological treatment. Depending on the progression of the disease, SGLT2 inhibitors, GLP-1 RA, DPP-4 inhibitors, sulphonylureas, thiazolidinediones, and insulin have also been recommended for glycemic control in T2DM patients (ADA 2020). Treatment for T2DM aims to reduce hyperglycemia by two main mechanisms: increased secretion of insulin by the pancreas or decreased production of glucose by the liver. However, T2DM is a progressive condition, that makes its treatment complicated, mainly due to the lack of control in insulin secretion and progressive cell death that leads to β-cell dysfunction, which should adjust the amount of insulin secreted in accordance with the needs of the organism. Therefore, patients often have episodes of hypoglycemia and hyperglycemia, both linked to serious complications of diabetes.

15.3.1 Oxidative Stress, Mitochondrial Dysfunction, and DNA Damage in T2DM Patients

In T2DM, hyperglycemia contributes significantly to the production of ROS, especially due to an overproduction of superoxide and hydrogen peroxide by the mitochondrial electron-transport chain (Dodson et al. 2013). Hyperglycemia can also induce the formation of AGEs, from protein glycation, which in turn contribute even more to ROS generation, thus aggravating the oxidative stress condition and leading to oxidative damage (Reddy et al. 2013; Pugazhenthi et al. 2017). In this context, several studies evaluating T2DM patients have shown increased levels of oxidative stress markers, measured by high levels of lipid peroxidation (thiobarbituric acid reactive substances (TBARS) and MDA), oxidized proteins, and AGEs (Abou-Seif and Youssef 2004; Strom et al. 2017) when compared to healthy individuals. The increased oxidative stress in T2DM patients is accompanied by decreased levels of antioxidant enzymes (GSH, SOD, and CAT) and total antioxidant status (Abou-Seif and Youssef 2004; Jiménez-Osorio et al. 2014; Strom et al. 2017). NRF2, a key protein involved in the transcription of genes belonging to the antioxidant response system, was also found in decreased levels in T2DM patients compared to healthy individuals (Jiménez-Osorio et al. 2014; Sireesh et al. 2018). Actually, newly diagnosed T2DM patients display decreased *NRF2* mRNA expression levels and reduced levels of its downstream target genes (SOD, HO-1, GPx, and CAT), increased mRNA expression levels of oxidative stress markers (p22Phox, TRPC6, and SOCS3), and increased levels of inflammatory cytokines (IL-4, IL-10, IL-13, IFN-γ, TNF-α, and GM-CSF). The reduced levels of NRF2, and consequently, low efficiency of the antioxidant response observed in T2DM patients can further aggravate the oxidative stress condition, contributing to the development of diabetes complications (Jiménez-Osorio et al. 2014). For instance, poor renal function in T2DM patients is associated with increased levels of lipid peroxidation (8-iso-PGF2α and MDA) (Sauriasari et al. 2015).

For this purpose, Golpour et al. (2020) performed a double-blind randomized placebo-controlled clinical trial and observed that a 10-week supplementation with fish oil n-3 PUFAs containing eicosapentaenoic (EPA) and docosahexaenoic (DHA) acids increased *NRF2* gene expression, as well as the total antioxidant status, and decreased lipid peroxidation (MDA) in T2DM patients compared to the placebo group. Another randomized clinical trial has shown that an eight-week supplementation with resveratrol increased the expression of both *NRF2* and *SOD* genes, increasing the total antioxidant capacity and decreasing the levels of carbonylated proteins, besides significantly reducing weight, BMI, and blood pressure levels in

T2DM patients, compared to those who did not receive any supplementation (Seyyedebrahimi et al. 2018). Thus, NRF2 upregulation additionally with the reduction of oxidative stress markers may have beneficial effects for T2DM patients.

It has been reported that T2DM patients show increased levels of oxidized bases in DNA (urinary 8-OHdG) (Tatsch et al. 2015) and in the nucleotide pool (serum 8-oxodG) (Sun et al. 2015), compared to healthy individuals, and this can also be a consequence of oxidative stress. Furthermore, high levels of 8-OHdG in T2DM patients have been accompanied by high levels of proinflammatory cytokines and higher insulin resistance, suggesting a relationship between inflammation, insulin resistance, and oxidative-induced damage in T2DM (Tatsch et al. 2015).

Concerning the importance of glycemic control, higher levels of protein oxidation and lipid peroxidation, and decreased antioxidant status are reported in hyperglycemic T2DM patients in comparison with non-hyperglycemic T2DM group of patients (Çakatay 2005; Lodovici et al. 2008; Bigagli et al. 2012). The high level of protein oxidation in T2DM patients without any comorbidity indicates that oxidative stress is related to hyperglycemia and may not be exclusively a consequence of complications of the disease. Besides, it is well known that oxidative stress induces DNA damage. Studies with T2DM patients have reported that hyperglycemic patients exhibit high levels of DNA damage and oxidative DNA damage compared to those of non-hyperglycemic T2DM patients (Lodovici et al. 2008; Xavier et al. 2015). In this line, Xavier et al. (2014) have shown that a one-week intervention to control glucose levels is efficient to significantly reduce DNA damage levels in T2DM patients compared to healthy individuals. Since high DNA damage levels are associated with the development of diabetes complications (Giacco and Brownlee 2010; Kumar et al. 2020), it is plausible to suggest that proper glycemic control may delay the progression of the disease and later complications.

Moreover, besides the evidence that T2DM patients show higher DNA damage than healthy subjects, when cells from T2DM patients were in vitro exposed to mutagens, it was found a lower efficiency of DNA repair mechanisms (Blasiak et al. 2004; Merecz et al. 2015). Curiously, Merecz et al. (2015) showed that T2DM patients with polymorphisms in *APE1* gene (a key gene in BER) showed different DNA repair capacities, and higher DNA damage levels compared to those without the polymorphism.

Since cellular respiration in mitochondria makes this organelle the site of increased production of ROS inside the cell (Peoples et al. 2019), some studies have evaluated different mitochondrial parameters in patients with diabetes mellitus. Bhansali et al. (2017) have shown that T2DM patients present high levels of mitochondrial ROS and several mitochondrial alterations, such as membrane depolarization, reduced mass, and morphological alterations, all of them being indicative of mitochondrial dysfunction. They also showed a downregulation of both mRNA and proteins (PINK1, MFN2, NIX, PARKIN, and LC3-II) associated with mitophagy, suggesting that an impaired mitophagy favors the accumulation of dysfunctional mitochondria and increases ROS production. Additionally, RNA sequencing in blood cells of T2DM patients has shown a downregulation of several mitochondrial genes, such *MT-ATP6, MT-ND1, MT-ND2, MT-ND4, MT-ND4L, MT-ND5, MT-ND6*

(Ustinova et al. 2020; Lv et al. 2020) related to mitochondrial oxidative phosphorylation and mitochondrial energy transduction compared to healthy subjects. Furthermore, the quantification of mtDNA copy number (mtDNA-CN) has been explored as a marker to assess mitochondrial function, suggesting that a greater number of mtDNA-CN is related to a better mitochondrial function. Accordingly, it has been found that T2DM patients show lower mtDNA-CN when compared to healthy subjects (Cho et al. 2017; Constantin-Teodosiu et al. 2020; Latini et al. 2020; Fazzini et al. 2021; Memon et al. 2021) and Latini et al. (2020) have suggested that diabetes complications are also associated with lower mtDNA-CN.

Therefore, these studies suggest that hyperglycemia is an important factor involved in oxidative stress, oxidative damage, and mitochondrial dysfunction, indicating the requirement of proper control of blood glucose levels and ROS production, in an attempt to reduce their detrimental effects on different macromolecules, such as nucleic acids, lipids and proteins, and avoiding later diabetes complications.

15.3.2 Transcriptional Gene Expression Profiles in T2DM and Alterations in Oxidative Stress and DNA Repair Genes

Despite the number of studies regarding T2DM, the molecular mechanisms involved in the development and progression of the disease still requires elucidation. In the last years, many studies have used large-scale transcriptomic analysis (microarray, RNA sequencing, single-cell RNA sequencing) to analyze mRNA, microRNA, long noncoding RNAs, and circulatory RNAs expression profiles exhibited by T2DM patients to reveal the main genes and pathways associated with the pathophysiological changes described in T2DM. Manoel-Caetano et al. (2012) conducted a study comparing the transcriptional expression patterns exhibited by PBMCs from T2DM patients compared with healthy subjects. The authors obtained a list of 92 differentially expressed genes (52 upregulated and 40 downregulated) in diabetic patients compared to the control group; among them, genes related to oxidative stress responses and hypoxia (*OXR1*, *SMG1*, and *UCP3*) were highly upregulated, possibly in an attempt to deal with increased oxidative stress. Regarding the downregulated genes, many were involved in inflammation, immune response, and DNA repair (including *SUMO1*, *ATRX*, and *MORF4L2*). The downregulation of several DNA repair genes is in agreement with the decreased efficiency of DNA repair reported for T2DM patients (Blasiak et al. 2004; Merecz et al. 2015). A study performed by Xavier et al. (2015) compared the mRNA transcriptional expression profiles of PBMCs from hyperglycemic, non-hyperglycemic T2DM patients and healthy individuals. Among the results, they found 478 genes (261 upregulated and 217 downregulated) differentially expressed related to several processes including upregulation of the inflammatory response process and the regulation of DNA repair, downregulation of response to superoxide, and response to the ER stress for the comparison between the hyperglycemic and nonhyperglycemic T2DM patients. Xavier et al. (2015) also found several differentially expressed miRNAs, such as hsa-miR-186, hsa-miR-222, and hsa-miR-29b, when comparing T2DM patients *versus* healthy controls, and these miRNAs were related to the development of β-islets, cell cycle regulation, and insulin resistance, respectively. The authors further searched for possible interactions between the miRNAs and the differentially expressed mRNAs providing new information to the pathogenesis of T2DM and the importance of adequate glycemic control.

The skeletal muscle also represents an important target tissue in the study of T2DM pathogenesis. A large RNA-seq based transcriptome study of human skeletal muscle of T2DM patients reported a significant downregulation of key genes involved in insulin signaling (such as *MTOR*, *PIK3CA*, *MAPK9*, *SLC2A4*, *PPARA*, *IRS2*), which is suggestive of impaired insulin action; oxidative phosphorylation, indicative of mitochondrial dysfunction and related to increased ROS generation; ER protein processing, which may be associated with ER stress; and upregulation of genes related to apoptosis, TP53 signaling, TNF-receptor family members and NF-kB signaling, indicative of increased cell death and inflammation in T2DM (Wu et al. 2017).

Another approach to the comprehension of T2DM etiopathogenesis has been reported in β-cells of pancreatic islets from human donors by Marselli et al. (2020), using RNA sequencing; the authors showed that T2DM islets have several molecular changes regarding upregulation of ROS activity, intracellular calcium regulation, apoptotic pathways, and metabolism of FFAs, whereas processes related to the mitochondrial respiratory chain and translational control were downregulated compared to the islets of healthy donors. Accordingly, Lundberg et al. (2018) showed that T2DM islets displayed downregulation of genes related to mitochondrial function, while genes associated with oxidative stress and UPR were upregulated. Pancreatic β-cell death has been associated with ER stress and UPR response, due to the high insulin demand, which causes an increased dependence on ER functioning to ensure proper synthesis and insulin folding (Cao et al. 2020). Komura et al. (2010) detected elevated expression of ER stress markers, comparing the transcriptional expression profiles of PBMCs from T2DM patients *versus* healthy individuals. Furthermore, Iwasaki et al. (2014) provided evidence that ATF4 (a transcription factor activated after metabolic stresses, including ER stress) was activated by FFAs in macrophages. Back et al. (2009) showed that the absence of eIF2α phosphorylation (responsible for activating ATF4) in mice β-cells caused dysregulated proinsulin translation, increased oxidative damage, and defective ER trafficking of proteins and apoptosis. In this context, Lytrivi et al. (2020) provided a comprehensive perspective of transcriptional changes in β-cells induced by FFAs. They found changes in lipid metabolism, ER stress, cell cycle, oxidative stress, and cAMP/PKA signaling implicated in the lipotoxicity of β-cells. Similarly, Bikopoulos et al. (2008) showed that aside from pancreatic islets chronically exposed to FFAs having a significantly reduced glucose-stimulated insulin secretion and increased ROS generation, they also presented altered expression of 40 genes mainly related to the FFAs metabolism, inflammation, and also to antioxidant defense (which were upregulated), highlighting the importance of FFAs as risk factors for the development of T2DM.

Another advanced technology, the single-cell RNA sequencing (scRNA-seq), has been used to analyze transcriptional profiles of T2DM patients at the β-cell level. Bosi et al. (2020) have made an integrative analysis of large scRNA-seq studies of human islets to identify molecular alterations and provide new relevant information for T2DM pathogenesis. They identified 226 differentially expressed genes (210 upregulated and 16 downregulated) that included upregulation of pathways linked to lysosome activity and ER stress/UPR and downregulation of pathways associated with regulation of ROS, DNA repair, and also regulation of DNA damage checkpoint, among others. Curiously, 25 of the differentially expressed genes (mainly linked to β-cell damage, increased oxidative stress, ER stress, impaired insulin action, and autophagy) had not been previously associated with T2DM, which help to understand β-cell dysfunction in T2DM.

Regarding diabetes complications, Massaro et al. (2019) found several miRNAs associated with specific diabetes complications. The miR-144-3p, for example, was found differentially expressed in both T1DM and T2DM, and its targets are linked to impaired insulin signaling pathway (*IRS1, TGF-β1, and PTEN*). These findings strongly correlate with insulin resistance and diabetes development (White 2014). Moreover, the gene atlas reported 650 nonredundant genes related to specific complications of diabetes (Rani et al. 2017), and seven genes (*AGER*, *TNFRSF11B*, *CRK*, *PON1*, *CRP,* and *NOS3*) were reported to be associated with cardiovascular diseases, nephropathy, retinopathy, and neuropathy, which are complications of the disease. Furthermore, the authors also reported miRNAs associated with diabetes complications; the hsa-miR-107, for instance, common to all complications, is associated with ER stress-induced lipid accumulation. Other miRNAS, such as mir-802, mir-181, mir-34a, and mir-24a, have been suggested as novel potential therapeutic targets, associated with impaired glucose metabolism, insulin resistance, and β-cell death and dysfunction, respectively (Rani et al. 2017).

Taken together, information in the literature on transcriptional expression profiles highlights not only a serious picture of changes in different molecular signaling pathways in T2DM (such as inflammation, oxidative stress response, DNA repair, apoptosis, antioxidant response, mitochondrial function, immune response, and ER stress, among others) but also establishes a link and integration between biological processes, which are clearly related to the physiological changes presented by patients. In addition, the use of microarrays and other advanced techniques for the study of large-scale transcriptional profiles brings an immense amount of data, revealing altered pathways still unknown in T2DM, thus expanding knowledge about the disease and also providing valuable data for new therapeutic and diagnostic possibilities.

15.4 Conclusions

Diabetes mellitus is a worldwide public health problem characterized by disturbances in the control of glucose levels, generating hyperglycemia and serious consequences for the body. Several multidisciplinary studies have shown that hyperglycemia has a major impact on the onset of oxidative stress and mitochondrial dysfunction, which is strongly correlated with damage to β-cells and the progression of the disease toward various complications. There is evidence that in both types, T1DM and T2DM, there is an increase of oxidative stress (as it can be detected by various molecular and biochemical markers), oxidative damage to macromolecules, as well as decreased antioxidant response and impaired DNA repair capacity. Studies on a genomic scale have shown that patients with T1DM and T2DM present important transcriptional changes related to a series of biological processes, especially regarding responses to oxidative stress, DNA repair, inflammation, immune response, ER stress, and mitochondrial alterations, among other processes, which reflects the characteristic picture of physiological changes that occur during development and progression of the disease. Therefore, all these changes described for diabetes show the need for adequate control of glycemia, in an attempt to reduce the deleterious effects over the years of chronic disease, as well as to delay its progression and thus avoiding subsequent complications of diabetes.

References

Abdel-Moneim A, Zanaty MI, El-Sayed A et al (2020) Relation between oxidative stress and hematologic abnormalities in children with type 1 diabetes. Can J Diabetes 44:222–228

Abou-Seif MA, Youssef AA (2004) Evaluation of some biochemical changes in diabetic patients. Clin Chim Acta 346:161–170

ADA (2020) Standards of medical care in diabetes—2021 abridged for primary care providers. Clin Diabetes cd21as01

Agbu P, Carthew RW (2021) MicroRNA-mediated regulation of glucose and lipid metabolism. Nat Rev Mol Cell Biol:1–14

Ahmad W, Ijaz B, Shabbiri K et al (2017) Oxidative toxicity in diabetes and Alzheimer's disease: mechanisms behind ROS/ RNS generation. J Biomed Sci 24:1–10

Akash MSH, Rehman K, Chen S (2013) Role of inflammatory mechanisms in pathogenesis of type 2 diabetes mellitus. J Cell Biochem 114:525–531

Altincik A, Tuğlu B, Demir K et al (2016) Relationship between oxidative stress and blood glucose fluctuations evaluated with daily glucose monitoring in children with type 1 diabetes mellitus. J Pediatr Endocrinol Metab 29:435–439

Assmann TS, Recamonde-Mendoza M, De Souza BM, Crispim D (2017) MicroRNA expression profiles and type 1 diabetes mellitus: systematic review and bioinformatic analysis. Endocr Connect 6:773–790

Ates I, Kaplan M, Yuksel M et al (2016) Determination of thiol/disulphide homeostasis in type 1 diabetes mellitus and the factors associated with thiol oxidation. Endocrine 51:47–51

Back SH, Scheuner D, Han J et al (2009) Translation attenuation through eIF2α phosphorylation prevents oxidative stress and maintains the differentiated state in β cells. Cell Metab 10:13–26

Balzano-Nogueira L, Ramirez R, Zamkovaya T et al (2021) Integrative analyses of TEDDY Omics data reveal lipid metabolism abnormalities, increased intracellular ROS and heightened inflammation prior to autoimmunity for type 1 diabetes. Genome Biol 22:39

Basile KJ, Johnson ME, Xia Q, Grant SFA (2014) Genetic susceptibility to type 2 diabetes and obesity: Follow-up of findings from genome-wide association studies. Int J Endocrinol 2014:769671

Bhansali S, Bhansali A, Walia R et al (2017) Alterations in mitochondrial oxidative stress and mitophagy in subjects with prediabetes and type 2 diabetes mellitus. Front Endocrinol (Lausanne) 8:347

Bigagli E, Raimondi L, Mannucci E et al (2012) Lipid and protein oxidation products, antioxidant status and vascular complications in poorly controlled type 2 diabetes. Br J Diabetes Vasc Dis 12:33–39

Bikopoulos G, da Silva PA, Lee SC et al (2008) Ex vivo transcriptional profiling of human pancreatic islets following chronic exposure to monounsaturated fatty acids. J Endocrinol 196:455–464

Bjørklund G, Dadar M, Pivina L et al (2019) The role of zinc and copper in insulin resistance and diabetes mellitus. Curr Med Chem 27:6643–6657

Blasiak J, Arabski M, Krupa R et al (2004) DNA damage and repair in type 2 diabetes mellitus. Mutat Res Mol Mech Mutagen 554:297–304

Bosi E, Marselli L, De Luca C et al (2020) Integration of single-cell datasets reveals novel transcriptomic signatures of-cells in human type 2 diabetes. NAR Genomics Bioinforma 2:lqaa097

Boucher J, Kleinridders A, Kahn CR (2014) Insulin receptor signaling in normal. Cold Spring Harb Perspect Biol 6:a009191

Cadet J, Davies KJA (2017) Oxidative DNA damage & repair: an introduction. Free Radic Biol Med 107:2–12

Çakatay U (2005) Protein oxidation parameters in type 2 diabetic patients with good and poor glycaemic control. Diabetes Metab 31:551–557

Cao ZH, Wu Z, Hu C et al (2020) Endoplasmic reticulum stress and destruction of pancreatic β cells in type 1 diabetes. Chin Med J 133:68–73

Caramori ML, Kim Y, Goldfine AB et al (2015) Differential gene expression in diabetic nephropathy in individuals with type 1 diabetes. J Clin Endocrinol Metab 100:E876–E882

Chatterjee S, Mudher A (2018) Alzheimer's disease and type 2 diabetes: a critical assessment of the shared pathological traits. Front Neurosci 12:383

Chen Z, Yu R, Xiong Y et al (2017) A vicious circle between insulin resistance and inflammation in nonalcoholic fatty liver disease. Lipids Health Dis 16:1–9

Cho SB, Koh I, Nam HY et al (2017) Mitochondrial DNA copy number augments performance of A1C and oral glucose tolerance testing in the prediction of type 2 diabetes. Sci Rep 7:43203

Codoñer-Franch P, Pons-Morales S, Boix-García L, Valls-Bellés V (2010) Oxidant/antioxidant status in obese children compared to pediatric patients with type 1 diabetes mellitus. Pediatr Diabetes 11:251–257

Collin F (2019) Chemical basis of reactive oxygen species reactivity and involvement in neurodegenerative diseases. Int J Mol Sci 20:2407

Constantin-Teodosiu D, Constantin D, Pelsers MM et al (2020) Mitochondrial DNA copy number associates with insulin sensitivity and aerobic capacity, and differs between sedentary, overweight middle-aged males with and without type 2 diabetes. Int J Obes 44:929–936

Cui H, Kong Y, Zhang H (2012) Oxidative stress, mitochondrial dysfunction, and aging. J Signal Transduct 2012:1–13

de Souza-Pinto NC, Maynard S, Hashiguchi K et al (2009) The recombination protein RAD52 cooperates with the excision repair protein OGG1 for the repair of oxidative lesions in mammalian cells. Mol Cell Biol 29:4441–4454

DeFronzo RA, Ferrannini E, Groop L et al (2015) Type 2 diabetes mellitus. Nat Rev Dis Prim 1:15019

Delbarba A, Abate G, Prandelli C et al (2016) Mitochondrial alterations in peripheral mononuclear blood cells from Alzheimer's disease and mild cognitive impairment patients. Oxidative Med Cell Longev 2016:5923938

Devlin C, Greco S, Martelli F, Ivan M (2011) MiR-210: more than a silent player in hypoxia. IUBMB Life 63:94–100

DiMeglio LA, Evans-Molina C, Oram RA (2018) Type 1 diabetes. Lancet 391:2449–2462

Dinçer Y, Akçay T, Ilkova H et al (2003) DNA damage and antioxidant defense in peripheral leukocytes of patients with Type I diabetes mellitus. Mutat Res - Fundam Mol Mech Mutagen 527:49–55

Dizdaroglu M, Coskun E, Jaruga P (2017) Repair of oxidatively induced DNA damage by DNA glycosylases: mechanisms of action, substrate specificities and excision kinetics. Mutat Res - Rev Mutat Res 771:99–127

Djeli N, Radakovi M, Dimirijevi V et al (2019) Oxidative stress and DNA damage in peripheral blood mononuclear cells from normal, obese, prediabetic and diabetic persons exposed to adrenaline in vitro. Mutat Res Toxicol Environ Mutagen 843:81–89

Dodson M, Darley-Usmar V, Zhang J (2013) Cellular metabolic and autophagic pathways: Traffic control by redox signaling. Free Radic Biol Med 63:207–221

El Amrousy D, El-Afify D, Shabana A (2021) Relationship between bone turnover markers and oxidative stress in children with type 1 diabetes mellitus. Pediatr Res 89:878–881

Fazzini F, Lamina C, Raftopoulou A et al (2021) Association of mitochondrial DNA copy number with metabolic syndrome and type 2 diabetes in 14176 individuals. J Intern Med 290:190–202

Gastol J, Polus A, Biela M et al (2020) Specific gene expression in type 1 diabetic patients with and without cardiac autonomic neuropathy. Sci Rep 10:1–8

Gerber PA, Rutter GA (2017) The role of oxidative stress and hypoxia in pancreatic beta-cell dysfunction in diabetes mellitus. Antioxidants Redox Signal 26:501–518

Gheni DA, Al-Maamori JA, Ghali KH (2020) The impact of oxidative stress and some endogenous antioxidants on type 1 diabetes mellitus. Eur J Mol Clin Med 7:4295–4310

Giacco F, Brownlee M (2010) Oxidative stress and diabetic complications. Circ Res 107:1058–1070

Golpour P, Nourbakhsh M, Mazaherioun M et al (2020) Improvement of NRF2 gene expression and antioxidant status in patients with type 2 diabetes mellitus after supplementation with omega-3 polyunsaturated fatty acids: A double-blind randomised placebo-controlled clinical trial. Diabetes Res Clin Pract 162:108120

Gonzalez-Martin A, Adams BD, Lai M et al (2016) The microRNA miR-148a functions as a critical regulator of B cell tolerance and autoimmunity. Nat Immunol 17:433–440

Goodarzi MT, Navidi AA, Rezaei M, Babahmadi-Rezaei H (2010) Oxidative damage to DNA and lipids: correlation with protein glycation in patients with type 1 diabetes. J Clin Lab Anal 24:72–76

Hanawalt PC, Wilson SH (2016) Cutting-edge Perspectives in Genomic Maintenance III: Preface. DNA Repair (Amst) 44:1–3

Hannon-Fletcher MPA, O'Kane MJ, Moles KW et al (2000) Levels of peripheral blood cell DNA damage in insulin dependent diabetes mellitus human subjects. Mutat Res - DNA Repair 460:53–60

He F, Ru X, Wen T (2020) NRF2, a transcription factor for stress response and beyond. Int J Mol Sci 21:1–23

He L, He T, Farrar S et al (2017) Antioxidants maintain cellular redox homeostasis by elimination of reactive oxygen species. Cell Physiol Biochem 44:532–553

Hegde ML, Izumi T, Mitra S (2012) Oxidized base damage and single-strand break repair in mammalian genomes: Role of disordered regions and posttranslational modifications in early enzymes. Prog Mol Biol Transl Sci 110:123–153

Homuth G, Wahl S, Müller C et al (2015) Extensive alterations of the whole-blood transcriptome are associated with body mass index: Results of an mRNA profiling study involving two large population-based cohorts. BMC Med Genet 8:65

Huang X, Liu G, Guo J, Su ZQ (2018) The PI3K/AKT pathway in obesity and type 2 diabetes. Int J Biol Sci 14:1483–1496

Hurrle S, Hsu WH (2017) The etiology of oxidative stress in insulin resistance. Biom J 40:257–262

International Diabetes Federation (2019) IDF Diabetes Atlas, 9th edn. International Diabetes Federation, Brussels

Irvine KM, Gallego P, An X et al (2012) Peripheral blood monocyte gene expression profile clinically stratifies patients with recent-onset type 1 diabetes. Diabetes 61:1281–1290

Iwasaki Y, Suganami T, Hachiya R et al (2014) Activating transcription factor 4 links metabolic stress to interleukin-6 expression in macrophages. Diabetes 63:152–161

Jiménez-Osorio AS, Picazo A, González-Reyes S et al (2014) Nrf2 and redox status in prediabetic and diabetic patients. Int J Mol Sci 15:20290–20305

Kalyanaraman B (2013) Teaching the basics of redox biology to medical and graduate students: Oxidants, antioxidants and disease mechanisms. Redox Biol 1:244–257

Katsarou A, Gudbjörnsdottir S, Rawshani A et al (2017) Type 1 diabetes mellitus. Nat Rev Dis Prim 3:1–17

Komura T, Sakai Y, Honda M et al (2010) CD14+ monocytes are vulnerable and functionally impaired under endoplasmic reticulum stress in patients with type 2 diabetes. Diabetes 59:634–643

Krokan HE, Bjørås M (2013) Base excision repair. Cold Spring Harb Perspect Biol 5:1–22

Kumar V, Agrawal R, Pandey A et al (2020) Compromised DNA repair is responsible for diabetes-associated fibrosis. EMBO J 39:e103477

Kwak SH, Park KS (2016) Recent progress in genetic and epigenetic research on type 2 diabetes. Exp Mol Med 48:e220

Latini A, Borgiani P, De Benedittis G et al (2020) Mitochondrial DNA copy number in peripheral blood is reduced in type 2 diabetes patients with polyneuropathy and associated with a MIR499A gene polymorphism. DNA Cell Biol 39:1467–1472

Lin CC, Huang HH, Hu CW et al (2014) Trace elements, oxidative stress and glycemic control in young people with type 1 diabetes mellitus. J Trace Elem Med Biol 28:18–22

Lodovici M, Giovannelli L, Pitozzi V et al (2008) Oxidative DNA damage and plasma antioxidant capacity in type 2 diabetic patients with good and poor glycaemic control. Mutat Res - Fundam Mol Mech Mutagen 638:98–102

Lundberg M, Stenwall A, Tegehall A et al (2018) Expression profiles of stress-related genes in islets from donors with progressively impaired glucose metabolism. Islets 10:69–79

Luo J, Mills K, le Cessie S et al (2020) Ageing, age-related diseases and oxidative stress: what to do next? Ageing Res Rev 57:100982

Lv B, Bao X, Li P et al (2020) Transcriptome sequencing analysis of peripheral blood of type 2 diabetes mellitus patients with thirst and fatigue. Front Endocrinol (Lausanne) 11:558344

Lytrivi M, Ghaddar K, Lopes M et al (2020) Combined transcriptome and proteome profiling of the pancreatic β-cell response to palmitate unveils key pathways of β-cell lipotoxicity. BMC Genomics 21:590

Manoel-Caetano FS, Xavier DJ, Evangelista AF et al (2012) Gene expression profiles displayed by peripheral blood mononuclear cells from patients with type 2 diabetes mellitus focusing on biological processes implicated on the pathogenesis of the disease. Gene 511:151–160

Marselli L, Piron A, Suleiman M et al (2020) Persistent or transient human β cell dysfunction induced by metabolic stress: specific signatures and shared gene expression with type 2 diabetes. Cell Rep 33:108466

Massaro JD, Polli CD, Silva MCE et al (2019) Post-transcriptional markers associated with clinical complications in Type 1 and Type 2 diabetes mellitus. Mol Cell Endocrinol 490:1–14

Memon AA, Sundquist J, Hedelius A et al (2021) Association of mitochondrial DNA copy number with prevalent and incident type 2 diabetes in women: A population-based follow-up study. Sci Rep 11:4608

Meng X, Gong C, Cao B et al (2015) Glucose fluctuations in association with oxidative stress among children with T1DM: comparison of different phases. J Clin Endocrinol Metab 100:1828–1836

Merecz A, Markiewicz L, Sliwinska A et al (2015) Analysis of oxidative DNA damage and its repair in Polish patients with diabetes mellitus type 2: Role in pathogenesis of diabetic neuropathy. Adv Med Sci 60:220–230

Mittal K, Katare DP (2016) Shared links between type 2 diabetes mellitus and Alzheimer's disease: A review. Diabetes Metab Syndr Clin Res Rev 10:S144–S149

Moldogazieva NT, Mokhosoev IM, Mel'Nikova TI et al (2019) Oxidative stress and advanced lipoxidation and glycation end products (ALEs and AGEs) in aging and age-related diseases. Oxidative Med Cell Longev 2019:3085756

Nassima M, Djamila AA, Baya G, Hafida M (2014) Oxidative stress biomarkers during Type 1 diabetes in Algerian children. Clin Biochem 47:776–777

Newsholme P, Cruzat VF, Keane KN et al (2016) Molecular mechanisms of ROS production and oxidative stress in diabetes. Biochem J 473:4527–4550

Nyaga DM, Vickers MH, Jefferies C et al (2018) The genetic architecture of type 1 diabetes mellitus. Mol Cell Endocrinol 477:70–80

Pastore A, Ciampalini P, Tozzi G et al (2012) All glutathione forms are depleted in blood of obese and type 1 diabetic children. Pediatr Diabetes 13:272–277

Peoples JN, Saraf A, Ghazal N et al (2019) Mitochondrial dysfunction and oxidative stress in heart disease. Exp Mol Med 51:1–13

Pugazhenthi S, Qin L, Reddy PH (2017) Common neurodegenerative pathways in obesity, diabetes, and Alzheimer's disease. Biochim Biophys Acta Mol basis Dis 1863:1037–1045

Rani J, Mittal I, Pramanik A et al (2017) T2DiACoD: a gene atlas of type 2 diabetes mellitus associated complex disorders. Sci Rep 7:6892

Reddy VP, Perry G, Cooke MS et al (2013) Mechanisms of DNA damage and repair in Alzheimer disease. In: Madame curie bioscience database. Landes Bioscience, Austin

Rodrigues R, de Medeiros LA, Cunha LM et al (2018) Correlations of the glycemic variability with oxidative stress and erythrocytes membrane stability in patients with type 1 diabetes under intensive treatment. Diabetes Res Clin Pract 144:153–160

Ruia S, Saxena S, Prasad S et al (2016) Correlation of biomarkers thiobarbituric acid reactive substance, nitric oxide and central subfield and cube average thickness in diabetic retinopathy: a cross-sectional study. Int J Retin Vitr 2:8

Salmonowicz B, Krzystek-Korpacka M, Noczyńska A (2014) Trace elements, magnesium, and the efficacy of antioxidant systems in children with type 1 diabetes mellitus and in their siblings. Adv Clin Exp Med 23:259–268

Sauriasari R, Andrajati R, Azizahwati et al (2015) Marker of lipid peroxidation related to diabetic nephropathy in Indonesian type 2 diabetes mellitus patients. Diabetes Res Clin Pract 108:193–200

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Seyyedebrahimi S, Khodabandehloo · Hadi, Ensieh ·, et al (2018) The effects of resveratrol on markers of oxidative stress in patients with type 2 diabetes: a randomized, double-blind, placebo-controlled clinical trial. Acta Diabetol 55:341–353

Sheedy FJ, Palsson-Mcdermott E, Hennessy EJ et al (2010) Negative regulation of TLR4 via targeting of the proinflammatory tumor suppressor PDCD4 by the microRNA miR-21. Nat Immunol 11:141–147

Sies H, Jones DP (2020) Reactive oxygen species (ROS) as pleiotropic physiological signalling agents. Nat Rev Mol Cell Biol 21:363–383

Sifuentes-Franco S, Pacheco-Moisés FP, Rodríguez-Carrizalez AD, Miranda-Díaz AG (2017) The role of oxidative stress, mitochondrial function, and autophagy in diabetic polyneuropathy. J Diabetes Res 2017:1673081

Sireesh D, Dhamodharan U, Ezhilarasi K et al (2018) Association of NF-E2 Related Factor 2 (Nrf2) and inflammatory cytokines in recent onset Type 2 Diabetes Mellitus. Sci Rep 8:1–10

Slupphaug G (2003) The interacting pathways for prevention and repair of oxidative DNA damage. Mutat Res Mol Mech Mutagen 531:231–251

Stechova K, Kolar M, Blatny R et al (2012) Healthy first-degree relatives of patients with type 1 diabetes exhibit significant differences in basal gene expression pattern of immunocompetent cells compared to controls: Expression pattern as predeterminant of autoimmune diabetes. Scand J Immunol 75:210–219

Storr SJ, Woolston CM, Zhang Y, Martin SG (2013) Redox environment, free radical, and oxidative DNA damage. Antioxidants Redox Signal 18:2399–2408

Strom A, Kaul K, Brüggemann J et al (2017) Lower serum extracellular superoxide dismutase levels are associated with polyneuropathy in recent-onset diabetes. Exp Mol Med 49:e394

Sun J, Lou X, Wang H et al (2015) Serum 8-hydroxy-2′-deoxyguanosine (8-oxo-dG) levels are elevated in diabetes patients. Int J Diabetes Dev Ctries 35:368–373

Surova O, Zhivotovsky B (2013) Various modes of cell death induced by DNA damage. Oncogene 32:3789–3797

Svilar D, Goellner EM, Almeida KH, Sobol RW (2011) Base excision repair and lesion-dependent subpathways for repair of oxidative DNA damage. Antioxidants Redox Signal 14:2491–2507

Takahashi P, Xavier DJ, Evangelista AF et al (2014) MicroRNA expression profiling and functional annotation analysis of their targets in patients with type 1 diabetes mellitus. Gene 539:213–223

Tatsch E, Carvalho JAMD, Hausen BS et al (2015) Oxidative DNA damage is associated with inflammatory response, insulin resistance and microvascular complications in type 2 diabetes. Mutat Res - Fundam Mol Mech Mutagen 782:17–22

Tsalamandris S, Antonopoulos AS, Oikonomou E et al (2019) The role of inflammation in diabetes: Current concepts and future perspectives. Eur Cardiol Rev 14:50–59

Ustinova M, Ansone L, Silamikelis I et al (2020) Whole-blood transcriptome profiling reveals signatures of metformin and its therapeutic response. PLoS One 15:e0237400

van Oostrom O, de Kleijn DPV, Fledderus JO et al (2009) Folic acid supplementation normalizes the endothelial progenitor cell transcriptome of patients with type 1 diabetes: a case-control pilot study. Cardiovasc Diabetol 8:47

Volpe CMO, Villar-Delfino PH, Dos Anjos PMF, Nogueira-Machado JA (2018) Cellular death, reactive oxygen species (ROS) and diabetic complications review-article. Cell Death Dis 9:119

Vujkovic M, Keaton JM, Lynch JA et al (2020) Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nat Genet 52:680–691

Walter F, O'Brien A, Concannon CG et al (2018) ER stress signaling has an activating transcription factor 6 (ATF6)-dependent "off-switch". J Biol Chem 293:18270–18284

White MF (2014) IRS2 integrates insulin/IGF1 signalling with metabolism, neurodegeneration and longevity. Diabetes Obes Metab 16:4–15

Włodarczyk M, Nowicka G (2019) Obesity, DNA damage, and development of obesity-related diseases. Int J Mol Sci 20:1146

Wu C, Xu G, Tsai SYA et al (2017) Transcriptional profiles of type 2 diabetes in human skeletal muscle reveal insulin resistance, metabolic defects, apoptosis, and molecular signatures of immune activation in response to infections. Biochem Biophys Res Commun 482:282–288

Xavier DJ, Takahashi P, Evangelista AF et al (2015) Assessment of DNA damage and mRNA/miRNA transcriptional expression profiles in hyperglycemic versus non-hyperglycemic patients with type 2 diabetes mellitus. Mutat Res Mol Mech Mutagen 776:98–110

Xavier DJ, Takahashi P, Manoel-Caetano FS et al (2014) One-week intervention period led to improvements in glycemic control and reduction in DNA damage levels in patients with type 2 diabetes mellitus. Diabetes Res Clin Pract 105:356–363

Yaşar Durmuş S, Şahin NM, Ergin M et al (2019) How does thiol/disulfide homeostasis change in children with type 1 diabetes mellitus? Diabetes Res Clin Pract 149:64–68

Chapter 16
Large-Scale Gene Expression in Monogenic and Complex Genetic Diseases

Anette S. B. Wolff, Adam Handel, and Bergithe E. Oftedal

16.1 Introduction

Our genome holds around 20,000–25,000 genes that serve as blueprints for building our proteins. An understanding of how these genes are regulated and transcribed, and how the protein repertoire is generated and maintained are central to human biology, health, and the pathophysiology of the disease. Immune function is crucial to all aspects of health, and we have therefore chosen to focus on immunological disorders in this chapter.

Monogenic diseases are rare disorders caused by mutations in one gene while polygenic diseases often are common disorders with a complex genetic landscape related to polymorphisms at multiple *loci* across the genome. In this chapter, we will critically assess the contribution of large-scale gene expression analyses to both disease contexts. Monogenic disorders arise from mutations in genes encoding proteins vital for immune function and are excellent model diseases with extreme disease phenotypes that have provided crucial information for understanding basic immunological functions. Knockout animal models are the ideal platform to study these disorders, since modifying one genetic *locus* is far more amenable to genetic engineering than for multiple *loci*. However, these monogenic disorders are rare,

A. S. B. Wolff · B. E. Oftedal (✉)
Department of Clinical Science, University of Bergen, Bergen, Norway

KG Jebsen Center for Autoimmune Disorders, University of Bergen, Bergen, Norway

Department of Medicine, Haukeland University Hospital, Bergen, Norway
e-mail: bergithe.oftedal@uib.no

A. Handel
Developmental Immunology, MRC Weatherall Institute of Molecular Medicine,
University of Oxford, Oxford, UK

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_16

and so obtaining sufficient numbers of samples to ensure accurate translation of animal models to human disease can be challenging.

The reverse problem exists in polygenic autoimmune disorders, where it may be possible to include thousands of individuals in large-scale genomic research. However, the generation of animal models that effectively mirror the complex genetic background of polygenic autoimmune disease as well as potential environmental precipitants is extremely difficult. The use of large public biobanks integrating longitudinal sampling with clinical information provides a promising platform for retrospective studies identifying individuals at risk of developing disease and disentangling causal relationships between gene expression and disease pathophysiology.

Translational studies in specific disorders can help to understand other immune diseases that are more difficult to study in human patients. By focusing on a few apt examples, this chapter aims to provide an overview of how recent progress in transcriptomic analyses has advanced our understanding of gene function and immunobiology.

16.2 Monogenic Diseases

Multiple genes have been found to have important roles in immune function. Mutations in such genes cause immune deficiencies, often associated with autoimmune disease (Azizi et al. 2021). Monogenic diseases can be dominant or recessive, autosomal or sex-linked. Dominant inherited diseases occur when a variant in one copy of a gene is sufficient to result in a clinical phenotype, while in recessive inherited diseases both alleles of the gene have mutations (Fig. 16.1). Disease-causing mutations can be found on all chromosomes, including the X and Y chromosomes. Incomplete inheritance is mostly observed in diseases characterized by dominant inheritance patterns, meaning that the disease phenotype does not always segregate with the genetic change.

We will in this part focus on three examples of monogenic disorders; namely autoimmune polyendocrine syndrome type 1 (APS-1), monogenic severe combined immune deficiency (SCID), and the immunodysregulation polyendocrinopathy enteropathy X-linked (IPEX) syndrome. APS-1 is caused by mutations in the *autoimmune regulator* (*AIRE*) gene which can cause disease both with recessive (classic form) and dominant inheritance patterns while mutations in the *adenosine deaminase (ADA)* gene cause immune deficiency by autosomal recessive genetic inheritance. IPEX is an example of an X chromosome-linked monogenic disorder, resulting in gender-specific inheritance pattern and originating from mutations in the *transcription factor forkhead box P3 (FOXP3)* gene. An overview of monogenic autoimmune and immune deficiency diseases is given in Table 16.1.

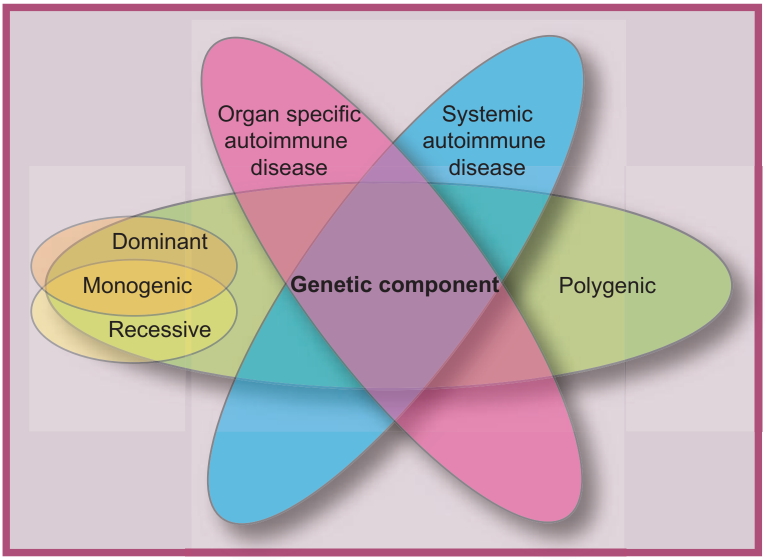

Organ specific autoimmune disease
Systemic autoimmune disease
Dominant
Monogenic
Recessive
Genetic component
Polygenic

Fig. 16.1 Overview of autoimmune disease and their patterns of inheritance. The genetic component of polygenic- and monogenic disorders, the latter with a dominant or recessive inheritance pattern is evident in both organ-specific and systemic autoimmune diseases

16.2.1 APS-1 and Central Tolerance Beyond AIRE

16.2.1.1 Clinical Phenotype and Molecular Genetics

APS-I is caused by autosomal inheritance of mutations in the *AIRE* gene on chromosome 21 (Finnish-German 1997; Nagamine et al. 1997). Although most commonly recessive in inheritance pattern, dominant mutations in *AIRE* causing a milder expression of the disease have also been described (Oftedal et al. 2015). The phenotype of APS-I is characterized by autoimmune endocrine manifestations; a clinical diagnosis requires two of Addison's disease, hypoparathyroidism, and chronic mucocutaneous candidiasis (CMC) (Husebye et al. 2018). Autoantibodies recognizing proteins expressed in the affected organs and against key intermediators of immune responses, notably type I interferons (IFNs) and interleukin (IL)-17 and -22, are hallmarks of this syndrome (Kisand et al. 2010; Puel et al. 2010).

Studying AIRE-deficient mice has provided the majority of the information regarding the etiology of APS-I with clear implications for human immune tolerance (Anderson et al. 2002; Ramsey et al. 2002; Liston et al. 2003, 2004; Anderson and Su 2016; Husebye et al. 2018). Mouse models have clear advantages in permitting one gene-one phenotype correlation. However, there are limitations to how accurately AIRE deficient mice model mimics human APS-1: mice show different autoimmune target organs (most commonly eyes, salivary glands, and exocrine pancreas), the mouse strain genetic background modulates disease severity, and the peripheral immune phenotype is different.

Table 16.1 Overview on monogenic autoimmune and immune deficiency disorders

Gene	Disease	Frequency	References
AIRE	Autoimmune polyendocrine syndrome type I (APS-I	1:100,000	Husebye et al. (2018)
BTK	X linked gammaglubolinemia (XLA), SCID	All SCID: 1/100,000	Vihinen et al. (2000)
ADA1	ADA-severe combined immune deficiency (SCID)	All SCID: 1/100,000	Fischer (2000)
ADA2 gene (formally CECR1)	deficiency of adenosine deaminase 2 (DADA2) (SCID)	All SCID: 1/100,000	Fischer (2000) and Kendall and Springer (2020)
IL2RG (*common gamma chain*)	X-SCID		Dvorak et al. (2019)
JAK3, CD45, Artemis, IL7Ra, CD3d, CD3E	SCID		Fischer (2000)
DCLREIC	Art-SCID		Schuetz et al. (2014)
FOXP3	X-linked Immune dysregulation polyendocrinopathy-enteropathy (IPEX)		Powell et al. (1982) and Park et al. (2020)
LRBA	Lipopolysaccharide responsive beige-like anchor (LRBA) deficiency, CVID		Lopez-Herrera et al. (2012)
TAC1, BAFF-B, MSHS, ….[a]	Common variable immunodeficiency (CVID)	1/10–50,000	Knight and Cunningham-Rundles (2006) and Ma et al. (2020)
gp91phox, p22phox, p47phox, p67phox, p40phox	Chronic granulomatous disease (CGD)		Arnold and Heimall (2017)
SH2D1A, XIAP	XLP	About 500 in total	Beenhouwer (2020)
WASp	Wiskott-Aldrich syndrome	1/250,000 males	Albert et al. (2011)
RAG1, RAG2	Omenn syndrome (SCID)		Villa et al. (1999)
ITGB2	Leukocyte adhesion deficiency (LAD1)		Fischer et al. (1988)
FAS, FASL	Autoimmune lymphoproliferative syndrome (ALPS)		Fisher et al. (1995)
Mutations in the complement system			Degn et al. (2011)

[a] Monogenic or "few genes" disorders in which mutations in one gene causes a disease phenotype

The main function of AIRE is to act as a transcriptional activator in medullary thymic epithelial cells (mTECs) during the process of thymic education of T cells (Anderson et al. 2002), although extrathymic expression of AIRE has been described (Gardner et al. 2008; Gardner et al. 2013). AIRE specifically acts in negative central tolerance, inducing the expression of ectopically proteins for presentation to developing T cells. Next generation sequencing has enabled whole transcriptomic studies to make major progress on understanding the function of AIRE in different thymic compartments.

16.2.1.2 Transcriptomics of AIRE Deficiency and Beyond

Focusing on AIRE's function in the thymus, mTECs can express nearly 20,000 unique proteins and AIRE has the potential to induce expression of up to 20% of these (Sansom et al. 2014). In 2015, Meredith and coworkers used single cell sequencing techniques to understand the function of AIRE and the mechanism of immune tolerance at the level of single mTECs. They performed parallel single-cell RNA-sequencing and DNA-single cell-methylation profiling of Aire wild type and Aire-deficient mTECs from mice (Meredith et al. 2015). Their results indicated that the organization of the DNA and epigenome is stochastically determined but preserved throughout cellular divisions. Holländer and coworkers showed that Aire-dependent genes in mTECs are marked with an epigenetic H3K27me3 repressive label, which allows them to be unsilenced epigenetically with subsequent expression (Sansom et al. 2014). To further delineate Aire's function, it was found that in addition to its role as a transcriptional activator, Aire has a repressive function to counteract accessibility of chromatins in tissue-specific gene loci (Koh et al. 2018). Hence, Aire calibrates the expression of tissue-specific genes by several mechanisms, and *AIRE* mutations impair both performances. These studies have made considerable progress regarding the function of Aire in the mouse thymus. However, as this disorder is very rare and thymic tissue inaccessible, transcriptomic studies in human patients have been more limited, with only a few studies published to date regarding the immune repertoire and transcriptome within specific immune subsets. There is surprising little effect on the global immune repertoire, although overall skewing of peripheral repertoire has been observed along with downregulation of peripheral FOXP3-positive regulatory T cells (Tregs) in APS-I (Kekalainen et al. 2007; Tuovinen et al. 2009; Laakso et al. 2010, 2011; Kaleviste et al. 2020).

Unlike most autoimmune disorders, the functional consequences of the loss of thymic AIRE expression in blood from APS-I-patients demonstrate repression of IFN and IFN-stimulated genes when the transcriptome is analyzed. This is probably due to the presence of high titer neutralizing anti-IFN antibodies (Kisand et al. 2008; Heikkila et al. 2016). Another study revealed impairments related to cell–cell signaling, innate immune responses, and cytokine activity in monocyte-derived dendritic cells from APS-I patients (Pontynen et al. 2008). A few studies have been published on TCR repertoires of different T cells subtypes regarding APS-I patients showing few differences between patients and controls (Niemi et al. 2015; Koivula et al. 2017), but Tregs from patients had a significantly longer TCR complementarity-determining region 3 than any other population, indicating that patients' naive Tregs have a defect already when entering the thymus (Koivula et al. 2017). The same group further showed that the CD8$^+$ T cell repertoire in APS-I patients was skewed compared to healthy controls (Laakso et al. 2011).

Understanding the mechanisms underlying human AIRE deficiency have been hampered by restricted accessibility to those tissues affected in APS-I. Recently, two different studies aimed to find the mechanisms underlying chronic candidiasis in APS-I patients by studying buccal biopsies with RNA sequencing. While Kaleviste and coworkers described an impaired antimicrobial response and cell proliferation profile characterizing APS-I patients, Break *et al*. found a localized increased IFN-γ and STAT1-response, which they hypothesized contributes to disruption of the epithelial membrane leading to infection (Kaleviste et al. 2020; Break et al. 2021), and the distinctive mucosal candidiasis seen in APS-I patients.

Intriguingly, common polymorphisms in the *AIRE* gene, combined with variants in other immune-related genes, are major risk factors of autoimmune Addison's disease in patients without APS-I (Eriksson et al. 2021). Hence, AIRE may have a role beyond monogenic APS-I in some polygenic autoimmune conditions which is yet to be functionally determined.

16.2.2 IPEX and Peripheral Tolerance

16.2.2.1 Clinical Phenotype and Molecular Genetics

The IPEX syndrome is a rare, X-linked disorder. IPEX usually presents in early childhood but may occur antenatally or later in life (Powell et al. 1982; Allenspach et al. 2017). It is characterized by severe enteropathy, chronic dermatitis, early onset type 1 diabetes mellitus (T1D), hypoparathyroidism, antibody-mediated cytopenia, and other autoimmune manifestations (Powell et al. 1982). Affected males typically die within the first years of life without immunosuppressive treatment or stem cell transplantation. IPEX was recognized clinically in 1982 (Powell et al. 1982), and linkage analysis mapped the genetic defect to the X chromosome (Xp11.23–Xq13.3) in 2000 (Bennett et al. 2000; Ferguson et al. 2000). Subsequently, the disease-causing variants were found to be in the *FOXP3* gene (a DNA-binding factor) when a frameshift mutation in *FOXP3* was observed in a naturally occurring mutant mouse (Bennett et al. 2001; Wildin et al. 2001).

The FoxP3 knock out mouse, also called the "scurfy mouse," shares many phenotypic features with the human disease, including scaly skin, low birth weight, diarrhea, progressive anemia, lymphadenopathy, and hepatosplenomegaly (Russell et al. 1959; Brunkow et al. 2001). The phenotype of the scurfy mouse results from immune dysregulation and loss of peripheral tolerance mechanisms due to uncontrolled proliferation of active CD4+ effector T cells (Blair et al. 1994; Clark et al. 1999), and an absence of regulatory T cells (Tregs) (Khattri et al. 2003).

As in the mouse, FOXP3 is the key factor in human Treg development, and its function has been extensively investigated (Ziegler 2006; Bin Dhuban and Piccirillo 2015; Alroqi and Chatila 2016; Bacchetta et al. 2018). Natural Tregs develop in the thymus, but there are also inducible Tregs (iTregs) in the periphery (Komatsu et al. 2009). When looking at these subtypes by single cell sequencing, their transcriptome of pTregs and iTregs were interlaced, but with minimal overlap in the T cell receptor repertoires (Hui et al. 2021). Tregs exert their suppressive effect either by direct cell–cell contact, secretion of immunoregulatory cytokines such as IL-10 and TGFβ (Palomares et al. 2014; Allenspach and Torgerson 2016), or by their high IL-2 affinity (Pandiyan et al. 2007).

16.2.2.2 Transcriptomics of FOXP3 Deficiency and Beyond

Previous studies have identified the existence of Treg-like cells in the absence of FOXP3 in some IPEX patients (Bacchetta et al. 2006; Otsubo et al. 2011; Boldt et al. 2014), and these cells have also been found in the Scurfy mice (Gavin et al. 2007; Lin et al. 2007; Charbonnier et al. 2019). Furthermore, the expression of FOXP3 has been found in cells that are otherwise similar to conventional T cells by single cell RNA sequencing (scRNAseq) (Zemmour et al. 2018). This, together with the finding of low expression-levels of FOXP3 early after the activation of conventional T cells (Walker et al. 2003; Gavin et al. 2006; Allan et al. 2007; McMurchy et al. 2013), suggest a role for FOXP3 outside the Treg compartment. Future studies will determine if this contributes to the skewing of effector T cell phenotypes described in IPEX (Passerini et al. 2011; Van Gool et al. 2019).

In a recent study by Zemmour et al. (2021), they included 15 IPEX patients and 15 healthy controls and characterized peripheral blood mononuclear cells (PBMCs) by scRNAseq, bulk RNA sequencing, and flow cytometry. Interestingly, they found the presence of heterogeneous Treg-like cells with an active FOXP3 locus in all patients. These Treg-like cells spanned a spectrum, where some resembled classical FOXP3-expressing Tregs, while others had distinct phenotypes. The dominant IPEX-signature found by Zemmour et al. was a monomorphic signature equally affecting all CD4+ T cells. Supported by the scurfy mouse model, the authors identified a cluster of genes that was regulated cell-intrinsically by FOXP3. Based on these findings, they suggested that FOXP3 is only important for very few Treg genes, including *IL-2ra*, *Tnfrsf4*, *Tnfrsf9*, *Tnfrsf18*, and *Capg*, corresponding to a "core set" of genes expressed by all Tregs, which are directly transactivated by and bind FOXP3 (Samstein et al. 2012; Kitagawa et al. 2017; Kwon et al. 2017). These genes encode the major homeostatic regulator of Treg cells (IL-2RA), and several members of the TNFR superfamily, which are also connected to Treg homeostasis and function (Chen et al. 2013; Remedios et al. 2018). This suggests a two-step model for the development of IPEX: cell-intrinsic downregulation of core FOXP3-responsive genes, which then drives global transcriptomic phenotypic differences in both Tregs and conventional T cells, and ultimately leads to a defective feedback cycle of T cell activation secondary to defective Treg function.

Transcriptional analyses have pinpointed the effect of FOXP3, and it will be interesting to see how this new information can be translated to new treatment options beyond IPEX and immunodeficiencies.

16.2.3 ADA Deficiency

16.2.3.1 Clinical Phenotype and Molecular Genetics

Monogenic mutations in the *ADA* gene encoding ADA1 result in the most common form of the SCID syndrome (Navon Elkan et al. 2014; Zhou et al. 2014; Kaljas et al. 2017) and clinically manifests as lymphopenia and immunodeficiency. The isoenzyme ADA2 is encoded by *ADA2/CECR1*. Both isoenzymes belong to the adenosine deaminase growth factor family and are important in the cellular metabolism through catalyzing the hydrolysis of adenosine to inosine in the purine catabolic pathway. ADA1 and -2 convert toxic deoxyadenosine, generated by DNA breakdown, to harmless metabolites. In addition to being a metabolic disorder, in a mouse model of ADA1 deficiency, the organs of the immune system including thymus and lymph nodes were decreased in size compared to healthy mice (Apasov et al. 2001). ADA1-SCID is inherited in an autosomal recessive manner and has an incidence of about 1:200,000, thus accounting for ~15% of all known SCID cases worldwide (Buckley 2004). ADA1 is expressed in most cells, whereas, in contrast, ADA2 is expressed almost exclusively in differentiating monocytes (Kaljas et al. 2017). High levels of ADA are expressed in lymphoid tissues with high cell turnover, especially in the thymus (Hirschhorn et al. 1978; Aldrich et al. 2003).

In ADA-deficient individuals there is a depletion of T, B, and NK cells, resulting in a failure to clear infections. Both ADA1 and ADA2 control the immune responses of nearby cells but acts on different targets; ADA1 connects via its receptor CD26 on the surface of effector T cells and NK cells and binds to CD16 negative monocytes while ADA2 anchors to neutrophils, monocytes, NK cells, B cells, and Tregs (Kaljas et al. 2017). ADA deficiency can be treated through enzyme replacement therapy with ADA-PEG injections, allogeneic hematopoietic stem cell transplantation, and hematopoietic stem cell gene therapy (Kohn et al. 2019).

16.2.3.2 Transcriptomics of ADA Deficiency

In both mice and humans with ADA1-deficiency, abnormal thymocyte development and differentiation is apparent (Apasov et al. 2001). The deleterious effect of mutations of ADA1 or -2 in T cells leads to the accumulation of toxic substrates not only within the cells but also through defective T cell signaling. A compromised TCR/CD28-driven proliferation and cytokine production has been revealed, associated with reduced ZAP-70 phosphorylation, Ca^{2+} flux, ERK1/2 signaling, and defective transcriptional events linked to CREB and NF-κB (Cassani et al. 2008). Flaws in B cells are probably caused by defect antigen-dependent maturation and compromised V(D)J recombination evident by reduced proliferation, increased apoptosis, and impaired signaling (Gangi-Peterson et al. 1999; Aldrich et al. 2003). When analyzing the transcriptome of monocytes from patients ADA2 deficiency in specific, increased levels of TNFα, IL6, and IL10 have been identified, in addition to upregulation of the IFN-type I pathway (Navon Elkan et al. 2014; Kaljas et al. 2017; Rama et al. 2018). Watanabe and colleagues demonstrated an expansion of CD16+ subtype and upregulation of both the IFNI and type II responses in monocytes from patients with ADA2 deficiency (Watanabe 2019). The global picture of the blood cells in ADA patients is still to be determined at single cell resolution at the transcriptional level.

16.3 Polygenic Diseases

Polygenic diseases are genetic disorders caused by the combined action or interaction of multiple genes (Fig. 16.1), often precipitated by unknown environmental triggers, exemplified by hormones, smoking, exposure to unknown chemical substances, and infections. In humans, these disorders occur more frequently than their monogenic counterparts, with a large social and economic impact. Some aspects of the genetic architecture of polygenic disorders are reasonably well understood (i.e., the implication of specific common single nucleotide polymorphisms) whereas the contributions of rare genetic variants and the mechanisms linking variants to disease pathogenesis are poorly understood. It is clear that polygenic analyses approaches and validation of genetic models need to be undertaken to understand the cumulative impact of the genes for immune disorders (Lvovs et al. 2012).

Mouse models of polygenic disorders are far more challenging than for the monogenic counterparts. Individual variants frequently have a minor impact on disease risk but engineering mice with the entire genetic background of polygenic risk variants is impractical. Environmental factors are important for the development of most polygenic conditions and these exposures are difficult to replicate in mouse colonies housed under sterile conditions.

For organ-specific disorders, advances have also been limited by the inaccessibility of some target organs for biopsy, the small size of the targeted organ, and because the target organ is frequently destroyed throughout the course of the disease.

In this section we will focus on the organ-specific polygenic diseases: type 1 diabetes, Graves' disease, vitiligo, and celiac disease; and the systemic polygenic autoimmune disorders: rheumatoid arthritis, systemic lupus erythematosus (SLE), and Sjøgren's syndrome.

16.3.1 Organ-Specific Disorders

16.3.1.1 Type 1 Diabetes

Type 1 diabetes (T1D) is caused by an autoimmune attack by T cells against the insulin-producing pancreatic β-cells, resulting in chronic hyperglycemia (Atkinson et al. 2014). The condition usually presents in childhood and is equally common in females and males. Adequate insulin replacement therapy can restore near-normal glycemic balance, although hyperglycemia might cause damage to other organs, including the retina, vascular structures, kidneys, and peripheral nerves. With a prevalence of ~1%, T1D is a relatively common organ-specific autoimmune disorder. Unlike APS-I and IPEX, the underlying genetic basis of T1D is thought to be polygenic, except for maturity-onset diabetes of the young (MODY), with contributions to disease susceptibility from poorly understood environmental factors.

The nonobese diabetic (NOD) mouse model is a promising tool to understand T1D pathogenesis (Giarratana et al. 2007). Indeed, the NOD mouse precipitates autoimmune glycemia, associates genetically to the orthologous human leukocyte antigen (HLA) region as in human T1D, and shows the same clinical picture and much of the pathogenic mechanisms as in the human corresponding disorder (Sharma et al. 2019). The prominent immune cell infiltration seen in NOD mice differs from the rather scarce infiltration seen in human insulitis (Campbell-Thompson et al. 2016). Single cell transcriptomics in NOD mice has demonstrated that the immune infiltration of the pancreas was highly dynamic (Zakharov et al. 2020). Until 2 weeks of age, there was little evidence of pancreatic infiltration but by 4 weeks memory CD4+ and CD8+ T cells predominated. Thereafter, Tregs, B cells, and dendritic cells entered the inflammatory milieu, in parallel with stepwise activation of macrophages.

A recent paper by Qian and colleagues took advantage of already published datasets on mRNA-expression data from the blood of diabetic patients and controls and performed global protein–protein interaction network analysis and transcription factor target network analysis based on differential expressions of genes (Qian et al. 2019). *CISC* was the most upregulated candidate, and *SCAF11* was the most downregulated molecule. In addition, they found *ARHGAP25, HLA-DRB4,* and *IL-23A* as promising and verified candidates. Lu and colleagues re-analyzed GWAS results for T1D combined with transcriptomic expressional data from the blood and revealed a network of genes responsible for immune regulation. They found several genes differentially expressed in T1D, including genes like *CAPZB, YWHAZ, TKT, TPP1, RBM17, PTPN11, HCG11,* and *MLLT1* (Lu et al. 2019). Other studies have been interested in detecting the autoreactive T cells in particular; peripheral T cells have been identified with TCRs specific for the common autoantibodies G6Pase2, insulin, pre-proinsulin, IA-2, GAD65, and ZnT8 in blood of T1D patients, but they are also seen in control subjects at low frequencies (Mallone et al. 2007; Eugster et al. 2015; Seay et al. 2016; Fuchs et al. 2017).

Studies of blood are convenient and may capture some of the important parameters to understand pathogenesis, however, the disease initiates at the surface of the pancreas and as the disease proceeds, within the organ. Studies in human pancreatic tissue from patients with T1D have confirmed that insulitis is associated with the presence of macrophages, NK cells, B cells, CD3+ T cells (CD8+ > CD4+), and HLA class I overexpression in the islets (Coppieters et al. 2012). There was a high degree of heterogeneity in T1D throughout the inflamed pancreas with T cell antigen reactivity, CD8+ T cell TCR clonality, and T cell phenotypes differing across individual islets (Coppieters et al. 2012; Arif et al. 2014; Pathiraja et al. 2015; Babon et al. 2016; Seay et al. 2016; Michels et al. 2017). This heterogeneity complicates inferences drawn from studies of peripheral blood. When analyzing 261 islets from either controls, islet-autoantibody-positive nondiabetic persons, and T1D patients, Campbell-Thompson et al. reported differential expression pathways related to β-cells and immune markers in antibody-positive nondiabetics, while β-cell transcripts were downregulated and HLA class II expression upregulated in T1D patients (Campbell-Thompson et al. 2018). Studying immune cells in different body compartments can gain further knowledge. In a recent study by Giola et al., they analyzed anti-islet T cells in the spleen and peripheral lymph nodes, the pancreatic lymph nodes, and within the islet cells in NOD mice (Gioia et al. 2019). In the periphery, the T cells had nearly no early T cell signaling genes expressed, although this was found both in pancreatic lymph nodes and intra-islet cells. Cytokine activation was detected in the pancreatic lymph node, however, no TCR recruitment was seen as opposed to the T cells within the target organ which had a profile of full TCR triggering and activation. The same study showed early epitope preference for a specific epitope within the insulin protein, which was not shown later, proving that autoimmunity toward insulin is part of the early event, and may be involved in triggering the loss of tolerance in these persons (Gioia et al. 2019).

The crosstalk between pancreatic epithelial cells and surface, and the surrounding immune cells are probably essential to initiate and maintain disease, and new technology opens new doors to investigate the locally affected milieu. Since pathogenic T cells are rare in the diseased pancreas (between 0.05% and 0.2%), and even rarer in the blood (<0.05% for CD4+ and <0.01% for CD8+ T cells) (Eugster et al. 2015), the combination of peptide–HLA–tetramers to enrich the pathogenic T cells for single cell RNA sequencing stands out as a powerful tool to understand the immunobiology and heterogeneity of islet-reactive T cells in T1D. This information can be applied to diagnose disease before the development of overt diabetes and to target disease-modifying treatments to specific islet-reactive clones (Zhang et al. 2021).

Complications of diabetes can further be explored using these novel approaches. Diabetic nephropathies confer damage to both the glomerulus and the tubule. To reveal changes in these regions, cells from the specified regions from cryopreserved kidneys were dissociated and single cell sequenced in three controls and compared to three diabetic biopsies (Wilson et al. 2019). Although there were not many leukocytes present, infiltrating monocytes expressed an IFN-γ and HLA class II profile. There was also an upregulation of B cells, T cells, plasma cells, and monocytes, and multiple cell types showed early signs of aberrant angiogenesis. Although unpowered, such molecular dissection of actual sites of pathogenic events may prove important to find new therapeutic tools, and within autoimmune diseases, T1D is one of the disorders where the analysis of the transcriptional landscape is most advanced.

16.3.1.2 Graves' Disease

Grave's disease (GD) is the most common autoimmune thyroid disease, with a lifetime risk of 3% for females and 0.5% for males. It is characterized by hyperthyroidism due to excessive production of thyroid hormone induced by TSH-receptor-specific stimulatory autoantibodies produced by B cells (Smith and Hegedus 2016). Autoreactive T cells infiltrate the tissue and produce cytokines, which maintain and amplify the immune response (Li et al. 2019). To date, GD is treated by controlling the unwanted production of thyroid hormone, by either antithyroid medications, radioactive iodine, or surgery. Much insight has been gained regarding the genetic susceptibility underlying Grave's disease, in which >20 genes contribute to the clinical phenotype, including immune regulatory genes (HLA, CTLA4, and PTPN22) and thyroid-specific genes (TG and TSHR). However, none of the known genetic variants contribute more than a fourfold increase of developing Graves' disease, with considerable heterogeneity between different populations (Davies et al. 2012).

Transcriptional analysis comparing thyroid tissue from patients with Graves' disease to control samples identified upregulation of human leukocyte antigen (HLA) genes, cytokine- and chemokine-related genes and regulators, and growth- and synthesis-related genes in Graves' disease (Yin et al. 2014). Pathway analysis identified five major categories of overactivity: immune responses (both innate and adaptive), pathogen-influenced gene expression, thyroid growth, stress responses, and intracellular and second-messenger signaling pathways. Some of these GD-related pathways may be driven by differential expression of specific microRNA molecules (Martinez-Hernandez et al. 2019). Similar, immune-dominant signatures define GD-associated inflammation outside of the thyroid. GD is frequently causing inflammation of extraocular adipose tissue, which has shown to be associated with IL-5 chemokine signaling by RNA sequencing (Lee et al. 2018). There is a rarity of studies in GD utilizing new transcriptome technology, where no single cell RNA sequencing data have been reported, either of GD thyroid samples or peripheral blood.

16.3.1.3 Vitiligo

Vitiligo is characterized by depigmented areas of the skin. The prevalence of this skin disorder worldwide is 0.2% but with geographical variation (Zhang et al. 2016). The pathogenic mechanisms causing vitiligo are thought to comprise autoimmunity against melanocytes, with subsequent reduction in melanin production (Katz and Harris 2021). CD8+ T cells and macrophages have been found within depigmented lesions in vitiligo, supporting an immune-mediated component in disease pathogenesis (van den Wijngaard et al. 2000).

Transcriptomic analyses have been performed on both blood and skin biopsies from patients with vitiligo and controls. In peripheral blood and affected skin, there was a common downregulation of immune/inflammatory responses, B cell pathways, apoptosis, and catabolic processes in addition to an altered interferon-profile involving *STAT1, STAT6, NFKB, CREB1,* and *MYC* amongst others (Dey-Rao and Sinha 2017). When analyzing skin samples, it is an advantage to compare lesioned skin with healthy skin from the same patient; such data show downregulated melanogenic pathways and of cornification and keratinocyte differentiation processes in damaged skin (Yu et al. 2012; Singh et al. 2017) and upregulation of innate immune system genes pointing at NK cell activity. Of interest, these immune genes were also shown to be upregulated in nonlesioned skin of vitiligo patients (Yu et al. 2012). Several studies have implicated interferons to be involved in pathogenic events in humans, although not in mice (Bertolotti et al. 2014; Rashighi et al. 2014; Tulic et al. 2019; Jacquemin et al. 2020; Riding et al. 2020). IL-15 is further increased in lesions (Chen et al. 2019) and the immune checkpoints CTLA4 and PD-1 are affected in blood and isolated T cells from vitiligo patients (Zhang et al. 2018; Willemsen et al. 2020), suggesting these pathways as suitable targets for developing and monitoring treatment of the patient.

16.3.1.4 Celiac Disease

Celiac disease is a systemic immune-mediated inflammation of the intestine triggered by dietary gluten. Celiac disease is characterized by a variety of clinical presentations, a specific serum autoantibody response, and damage to the small-intestinal mucosa (Fasano and Catassi 2012). The pathogenesis of this disease involves gluten as an external trigger, changes in intestinal permeability, enzymatically modified gluten, HLA recognition, and innate and adaptive immune responses to gluten peptides involving self-antigens (e.g., transglutaminase), eventually leading to celiac enteropathy (Jabri and Sollid 2009; Schuppan et al. 2009). A gluten-free diet is a principal therapy for celiac disease. The HLA haplotypes HLA-DQ2 and HLA-DQ8 are expressed in 90% and 5% of the patients, respectively. DQ2 and DQ8 haplotypes expressed on the surface of antigen-presenting cells bind activated (deamidated) gluten peptides, thereby triggering an abnormal immune response (Karell et al. 2003). In addition, genes involved in inflammatory and immune responses have been shown to predispose to celiac disease (Trynka et al. 2011).

Transcriptome analysis of FFPE archived duodenum biopsies from active celiac disease and control subjects demonstrated both dysregulation of genes associated with lymphocyte activation and cytokine response and impaired epithelial proliferation pathways in celiac disease (Loberman-Nachum et al. 2019). The integration of these data with two other RNA sequencing studies (Bragde et al. 2018; Leonard et al. 2019) identified a core celiac disease-specific signature of 403 genes.

The possibility to compare biologic material from the same patient with and without gluten challenge gives a remarkable opportunity to study the transcriptome in celiac disease and how it is affected by gluten. In a study by Donsenko et al. (2021), they found that healthy intestinal function could not be reinstalled upon long-term gluten-free diet restriction, and that gluten challenge induced hyperactive intestinal wnt signaling and consequent immature crypt gene expression resulting in a less differentiated epithelium. Investigation of the paired γδTCR repertoire found that celiac patients have a more diverse repertoire, but no specific γδTCR could be found that were specific for the patients, suggesting that identifying a celiac disease relevant γδTCR ligand to be difficult (Eggesbo et al. 2020).

16.3.2 Systemic

16.3.2.1 Rheumatoid Arthritis

Rheumatoid arthritis (RA) is an autoimmune disorder characterized by swollen, stiff, and painful joints. The cause is most likely to be multifactorial, including genetic predisposition, break of tolerance involving posttranslational modified proteins and environmental causes like, e.g., smoking (Sumitomo et al. 2018). Genetic susceptibility to RA predominantly maps to the HLA with other non-HLA contributions enriched within pathways associated with CD4+ T cell activation (Raychaudhuri et al. 2012; Ha et al. 2020).

Infiltrating B and T cells are found in synovial fluid and immune cells targeting citrullinated proteins in blood (Malmstrom et al. 2017), and upregulation of type I IFN signature genes in RA has been put forward, which is possible to target for therapy (He et al. 2008; Bremer et al. 2011; Sumitomo et al. 2018).

Previously, transcriptomic analyses were mostly done on blood samples from RA patients (Sanayama et al. 2014; Sellam et al. 2014), although the new era in genomics now facilitates analyses of pathological events near the joints and by using single cell approaches as a complement to bulk sequencing protocols. The inflamed joint is, however, a heterogenous and rather complex tissue with high dynamics and fast interactions with other cells and proteins. Sampling the affected joints, Carlberg and co-found an overabundance of T and B cells, especially T cells of the memory type at the localized site and claimed high levels of TNFα expression (Carlberg et al. 2019). A recent project described a systemic biology approach compiling data from single cell RNA transcriptomic, mass cytometry, bulk RNA seq, and flow cytometry of 51 synovial biopsies from RA patients, and showed several expanded cell subsets, including THY1(CD90)$^{+}$HLA-DRA^{hi} fibroblasts, IL-1B^{+} pro-inflammatory monocytes, ITGAX^{+}TBX21^{+} autoimmune-associated B cells and PDCD1^{+} peripheral helper T cells and follicular helper T cells. They further mapped aberrant IL-6 and IL-1B secretion to fibroblasts and monocytes and defined these as drivers of RA pathogenesis (Zhang et al. 2019a). By studying fibroblast-like synoviocytes, Galligans showed that 8 particular genes correlated with RA disease activity (Galligan et al. 2007); *(HLA)-DQA2, Clec12A, MAB21L2, SIAT7E, HAPLN1, BAIAP2L1, RGMB,* and *OSAP*. Other studies have been interested especially in determining positive outcomes of immune therapy; by examining involving synovial tissue biopsies, it was shown that myeloid, but not lymphoid, gene signature expression was associated with a positive effect of anti-TNF-treatment (Dennis et al. 2014).

Transcriptomic data on blood cells from RA patients found several links to aberrances in the type I interferon profile (Sumitomo et al. 2018), further used to monitor RA patients treated with immune therapy to predict outcome. Single cell analyses of whole blood by Sellam et al. revealed an upregulation of *NF-KB, IL-33*, and *STAT5A* and downregulation of the IFN pathway (Sellam et al. 2014). Koczan and colleagues predicted that TNFα signaling via *NFKB* could indicate a good or bad response toward TNF alpha therapy, including a gene set determined by *NFKBIA, CCL4, IL-8, IL-1B, TNFAIP3, PDE4B, PPP1R15A, and ADM* (Koczan et al. 2008). Others reported on dissection of blood CD4+ T cells, finding abnormal STAT1 and Wnt signaling (Ye et al. 2015), and in CD4+ subsets, showing upregulation of GTPases-associated signaling and apoptosis and dysregulation of the TCR pathway involving ZAP70 and JAK3 (He et al. 2008; Bremer et al. 2011; Sumitomo et al. 2018), molecules that can cause monogenic inborn errors of immunity if mutated on the germline level. A similar approach detected activation of IFN signaling in RA neutrophils and patients could be categorized in either IFN-high or IFN-low (Wright et al. 2015).

16.3.2.2 Systemic Lupus Erythematosus

Systemic lupus erythematosus (SLE) is a multisystem autoimmune disease with manifestations ranging from mild to severe and life-threatening. This is a challenging disease to manage due to its clinical heterogeneity combined with the potential severity of the symptoms (Liossis and Staveri 2021). SLE is a chronic disease that affects a variety of organ systems leading to organ failure due to the formation and deposition of autoantibodies and immune complexes (Rahman and Isenberg 2008). Despite recent advantages in treatment, SLE patients still have a two to threefold increased risk of death (Thomas et al. 2014; Yee et al. 2015). There is a genetic predisposition to SLE within families, especially within identical twins (Block et al. 1975; Deapen et al. 1992; Perdriger et al. 2003). Gene variants associated with SLE susceptibility are enriched within pathways regulating the function of multiple aspects of the innate and adaptive immune system (Tsokos 2011; Bentham et al. 2015).

Microarray-based transcriptomic analysis of SLE patients compared to healthy controls identified disease-associated transcriptional signatures related to type I interferon (IFN) and granulocytes (Baechler et al. 2003; Bennett et al. 2003). Single-cell RNA sequencing of renal samples from patients with lupus nephritis and control subjects confirmed global overactivation of type I IFN signaling Nature Immunol (2019). Twenty-one subsets of leukocytes were identified within lupus-associated renal tissue, including multiple populations of myeloid cells, T cells, natural killer cells, and B cells with both proinflammatory responses and inflammation-resolving responses. Trajectory analysis of monocytes within the inflamed kidney identified a continuum of intermediate states, including spanning patrolling, phagocytic, and alternatively activated cells. This suggests progressive stages of monocyte differentiation within the kidney to underpin lupus nephritis pathogenesis. Hence, access to the affected tissue and the use of single cell sequencing permit fine resolution understanding of the disease-associated immunobiology, emphasizing the potential of these new methods to be applied to other diseases.

16.3.2.3 Sjogren's Syndrome

Primary Sjögren's syndrome (pSS) is an autoimmune disorder presenting with dry mouth and eyes (xerostomia and xerophthalmia, respectively) (Manuel et al. 2017). The disease affects females more frequently than males with a ratio of 10:1. With the exception of HLA, risk genes (e.g., *STAT4, TNFAIP3, IL-12A, GTF2I, RBMS3*) have a minor effect on disease predisposition (Teos and Alevizos 2017).

Upregulation of antigen processing and presentation, humoral immune response, inflammatory reaction, and the Toll-like receptor and interferon pathways in the salivary glands have been identified as common pathways of disease pathogenesis both in mouse models and patient samples (Hjelmervik et al. 2005; Gottenberg et al. 2006; Horvath et al. 2012). Chemotaxis through the cytokines *CCR7* and *CCL21* were differentially expressed in salivary gland tissue from pSS patients relative to controls and were associated with *in vitro* evidence of Th17 predominance (Zhang et al. 2019b). Novel systemic approaches, joining genetic, transcriptional, proteomic, and clinical data, recently identified that *LINC00487* and *SOX4* expression were associated with B cell defects in pSS (Inamo et al. 2020). Seven "hub genes" were consistently differential expressed in multiple pSS transcriptomic datasets: *ICOS, SELL, CR2, BANK1, MS4A1, ZC3H12D*, and *CCR7*. *ICOS* was upregulated consistently in both salivary gland biopsies and blood, and furthermore found to be associated with lymphocytic infiltration and disease activity of pSS patients (Luo et al. 2020). Powerful studies like this can provide important information about pathogenesis and possible targets for therapy, by pinpointing consistently differentially regulated immune cell pathways and cellular subsets.

16.4 Conclusion

By focusing on diseases of the immune system, we have described the current use and the knowledge gained from large-scale gene expression studies in monogenic and complex disorders. While the monogenic studies often suffer from the limited number of patients included, the mouse models have proved important to elucidate the molecular function and the transcriptional consequences of the genetic abnormality. We found a large discrepancy in the utilization and progression into the new single cell transcriptomic analysis tools for the different diseases. While there were yet no single cell studies on GD, more severe diseases like SLE and T1D were highly represented, highlighting that there is still progress to be made in understanding the complete picture of these diseases and to define targets and pathways for future functional studies.

The still-emerging single cell technologies determining gene expression in cells and tissues will continue to yield important information in the years to come. Combined with algorithms incorporating genetic, lifestyle, and environmental factors, this holds the potential to further untangle these complex diseases.

References

Albert MH, Notarangelo LD, Ochs HD (2011) Clinical spectrum, pathophysiology and treatment of the Wiskott-Aldrich syndrome. Curr Opin Hematol 18(1):42–48

Aldrich MB, Chen W, Blackburn MR, Martinez-Valdez H, Datta SK, Kellems RE (2003) Impaired germinal center maturation in adenosine deaminase deficiency. J Immunol 171(10):5562–5570

Allan SE, Crome SQ, Crellin NK, Passerini L, Steiner TS, Bacchetta R, Roncarolo MG, Levings MK (2007) Activation-induced FOXP3 in human T effector cells does not suppress proliferation or cytokine production. Int Immunol 19(4):345–354

Allenspach E, Torgerson TR (2016) Autoimmunity and primary immunodeficiency disorders. J Clin Immunol 36(Suppl 1):57–67

Allenspach EJ, Finn LS, Rendi MH, Eken A, Singh AK, Oukka M, Taylor SD, Altman MC, Fligner CL, Ochs HD, Rawlings DJ, Torgerson TR (2017) Absence of functional fetal regulatory T cells in humans causes in utero organ-specific autoimmunity. J Allergy Clin Immunol 140(2):616–619 e617

Alroqi FJ, Chatila TA (2016) T regulatory cell biology in health and disease. Curr Allergy Asthma Rep 16(4):27

Anderson MS, Su MA (2016) AIRE expands: new roles in immune tolerance and beyond. Nat Rev Immunol 16(4):247–258

Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, Turley SJ, von Boehmer H, Bronson R, Dierich A, Benoist C, Mathis D (2002) Projection of an immunological self shadow within the thymus by the aire protein. Science 298(5597):1395–1401

Apasov SG, Blackburn MR, Kellems RE, Smith PT, Sitkovsky MV (2001) Adenosine deaminase deficiency increases thymic apoptosis and causes defective T cell receptor signaling. J Clin Invest 108(1):131–141

Arif S, Leete P, Nguyen V, Marks K, Nor NM, Estorninho M, Kronenberg-Versteeg D, Bingley PJ, Todd JA, Guy C, Dunger DB, Powrie J, Willcox A, Foulis AK, Richardson SJ, de Rinaldis E, Morgan NG, Lorenc A, Peakman M (2014) Blood and islet phenotypes indicate immunological heterogeneity in type 1 diabetes. Diabetes 63(11):3835–3845

Arnold DE, Heimall JR (2017) A review of chronic granulomatous disease. Adv Ther 34(12):2543–2557

Atkinson MA, Eisenbarth GS, Michels AW (2014) Type 1 diabetes. Lancet 383(9911):69–82

Azizi G, Tavakol M, Yazdani R, Delavari S, Moeini Shad T, Rasouli SE, Jamee M, Pashangzadeh S, Kalantari A, Shariat M, Shafiei A, Mohammadi J, Hassanpour G, Chavoshzadeh Z, Mahdaviani SA, Momen T, Behniafard N, Nabavi M, Bemanian MH, Arshi S, Molatefi R, Sherkat R, Shirkani A, Alyasin S, Jabbari-Azad F, Ghaffari J, Mesdaghi M, Ahanchian H, Khoshkhui M, Eslamian MH, Cheraghi T, Dabbaghzadeh A, Nasiri Kalmarzi R, Esmaeilzadeh H, Tafaroji J, Khalili A, Sadeghi-Shabestari M, Darougar S, Moghtaderi M, Ahmadiafshar A, Shakerian B, Heidarzadeh M, Ghalebaghi B, Fathi SM, Darabi B, Fallahpour M, Mohsenzadeh A, Ebrahimi S, Sharafian S, Vosughimotlagh A, Tafakoridelbari M, Rahimi Haji-Abadi M, Ashournia P, Razaghian A, Rezaei A, Salami F, Shirmast P, Bazargan N, Mamishi S, Ali Khazaei H, Negahdari B, Shokri S, Nabavizadeh SH, Bazregari S, Ghasemi R, Bayat S, Eshaghi H, Rezaei N, Abolhassani H, Aghamohammadi A (2021) Autoimmune manifestations among patients with monogenic inborn errors of immunity. Pediatr Allergy Immunol 32:1335–1348

Babon JA, DeNicola ME, Blodgett DM, Crevecoeur I, Buttrick TS, Maehr R, Bottino R, Naji A, Kaddis J, Elyaman W, James EA, Haliyur R, Brissova M, Overbergh L, Mathieu C, Delong T, Haskins K, Pugliese A, Campbell-Thompson M, Mathews C, Atkinson MA, Powers AC, Harlan DM, Kent SC (2016) Analysis of self-antigen specificity of islet-infiltrating T cells from human donors with type 1 diabetes. Nat Med 22(12):1482–1487

Bacchetta R, Passerini L, Gambineri E, Dai M, Allan SE, Perroni L, Dagna-Bricarelli F, Sartirana C, Matthes-Martin S, Lawitschka A, Azzari C, Ziegler SF, Levings MK, Roncarolo MG (2006) Defective regulatory and effector T cell functions in patients with FOXP3 mutations. J Clin Invest 116(6):1713–1722

Bacchetta R, Barzaghi F, Roncarolo MG (2018) From IPEX syndrome to FOXP3 mutation: a lesson on immune dysregulation. Ann N Y Acad Sci 1417(1):5–22

Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, Shark KB, Grande WJ, Hughes KM, Kapur V, Gregersen PK, Behrens TW (2003) Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc Natl Acad Sci U S A 100(5):2610–2615

Beenhouwer D (2020) Chapter 17 – Molecular basis of diseases of immunity. In: Coleman WB, Tsongalis GJ (eds) Essential concepts in molecular pathology. Academic Press, London, pp 271–284

Bennett CL, Yoshioka R, Kiyosawa H, Barker DF, Fain PR, Shigeoka AO, Chance PF (2000) X-Linked syndrome of polyendocrinopathy, immune dysfunction, and diarrhea maps to Xp11.23-Xq13.3. Am J Hum Genet 66(2):461–468

Bennett CL, Christie J, Ramsdell F, Brunkow ME, Ferguson PJ, Whitesell L, Kelly TE, Saulsbury FT, Chance PF, Ochs HD (2001) The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. Nat Genet 27(1):20–21

Bennett L, Palucka AK, Arce E, Cantrell V, Borvak J, Banchereau J, Pascual V (2003) Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med 197(6):711–723

Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, Martin J, Fairfax BP, Knight JC, Chen L, Replogle J, Syvanen AC, Ronnblom L, Graham RR, Wither JE, Rioux JD, Alarcon-Riquelme ME, Vyse TJ (2015) Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet 47(12):1457–1464

Bertolotti A, Boniface K, Vergier B, Mossalayi D, Taieb A, Ezzedine K, Seneschal J (2014) Type I interferon signature in the initiation of the immune response in vitiligo. Pigment Cell Melanoma Res 27(3):398–407

Bin Dhuban K, Piccirillo CA (2015) The immunological and genetic basis of immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome. Curr Opin Allergy Clin Immunol 15(6):525–532

Blair PJ, Bultman SJ, Haas JC, Rouse BT, Wilkinson JE, Godfrey VL (1994) CD4+CD8- T cells are the effector cells in disease pathogenesis in the scurfy (sf) mouse. J Immunol 153(8):3764–3774

Block SR, Winfield JB, Lockshin MD, D'Angelo WA, Christian CL (1975) Studies of twins with systemic lupus erythematosus. A review of the literature and presentation of 12 additional sets. Am J Med 59(4):533–552

Boldt A, Kentouche K, Fricke S, Borte S, Kahlenberg F, Sack U (2014) Differences in FOXP3 and CD127 expression in Treg-like cells in patients with IPEX syndrome. Clin Immunol 153(1):109–111

Bragde H, Jansson U, Fredrikson M, Grodzinsky E, Soderman J (2018) Celiac disease biomarkers identified by transcriptome analysis of small intestinal biopsies. Cell Mol Life Sci 75(23):4385–4401

Break TJ, Oikonomou V, Dutzan N, Desai JV, Swidergall M, Freiwald T, Chauss D, Harrison OJ, Alejo J, Williams DW, Pittaluga S, Lee CR, Bouladoux N, Swamydas M, Hoffman KW, Greenwell-Wild T, Bruno VM, Rosen LB, Lwin W, Renteria A, Pontejo SM, Shannon JP, Myles IA, Olbrich P, Ferre EMN, Schmitt M, Martin D, Genomics and Computational Biology Core, Barber DL, Solis NV, Notarangelo LD, Serreze DV, Matsumoto M, Hickman HD, Murphy

PM, Anderson MS, Lim JK, Holland SM, Filler SG, Afzali B, Belkaid Y, Moutsopoulos NM, Lionakis MS (2021) Aberrant type 1 immunity drives susceptibility to mucosal fungal infections. Science 371(6526):eaay5731

Bremer E, Abdulahad WH, de Bruyn M, Samplonius DF, Kallenberg CG, Armbrust W, Brouwers E, Wajant H, Helfrich W (2011) Selective elimination of pathogenic synovial fluid T-cells from rheumatoid arthritis and juvenile idiopathic arthritis by targeted activation of Fas-apoptotic signaling. Immunol Lett 138(2):161–168

Brunkow ME, Jeffery EW, Hjerrild KA, Paeper B, Clark LB, Yasayko SA, Wilkinson JE, Galas D, Ziegler SF, Ramsdell F (2001) Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse. Nat Genet 27(1):68–73

Buckley RH (2004) Molecular defects in human severe combined immunodeficiency and approaches to immune reconstitution. Annu Rev Immunol 22:625–655

Campbell-Thompson M, Fu A, Kaddis JS, Wasserfall C, Schatz DA, Pugliese A, Atkinson MA (2016) Insulitis and beta-cell mass in the natural history of type 1 diabetes. Diabetes 65(3):719–731

Campbell-Thompson M, Butterworth EA, Lenchik NI, Atkinson MA, Mathews CE, Gerling IC (2018) Analysis of transcriptome data from 261 individual laser-captured Islets from nondiabetic, autoantibody-positive, and Type 1 diabetic organ donors. Diabetes 67:308

Carlberg K, Korotkova M, Larsson L, Catrina AI, Stahl PL, Malmstrom V (2019) Exploring inflammatory signatures in arthritic joint biopsies with spatial transcriptomics. Sci Rep 9(1):18975

Cassani B, Mirolo M, Cattaneo F, Benninghoff U, Hershfield M, Carlucci F, Tabucchi A, Bordignon C, Roncarolo MG, Aiuti A (2008) Altered intracellular and extracellular signaling leads to impaired T-cell functions in ADA-SCID patients. Blood 111(8):4209–4219

Charbonnier LM, Cui Y, Stephen-Victor E, Harb H, Lopez D, Bleesing JJ, Garcia-Lloret MI, Chen K, Ozen A, Carmeliet P, Li MO, Pellegrini M, Chatila TA (2019) Functional reprogramming of regulatory T cells in the absence of Foxp3. Nat Immunol 20(9):1208–1219

Chen X, Wu X, Zhou Q, Howard OM, Netea MG, Oppenheim JJ (2013) TNFR2 is critical for the stabilization of the CD4+Foxp3+ regulatory T cell phenotype in the inflammatory environment. J Immunol 190(3):1076–1084

Chen X, Guo W, Chang Y, Chen J, Kang P, Yi X, Cui T, Guo S, Xiao Q, Jian Z, Li K, Gao T, Li S, Liu L, Li C (2019) Oxidative stress-induced IL-15 trans-presentation in keratinocytes contributes to CD8(+) T cells activation via JAK-STAT pathway in vitiligo. Free Radic Biol Med 139:80–91

Clark LB, Appleby MW, Brunkow ME, Wilkinson JE, Ziegler SF, Ramsdell F (1999) Cellular and molecular characterization of the scurfy mouse mutant. J Immunol 162(5):2546–2554

Coppieters KT, Dotta F, Amirian N, Campbell PD, Kay TW, Atkinson MA, Roep BO, von Herrath MG (2012) Demonstration of islet-autoreactive CD8 T cells in insulitic lesions from recent onset and long-term type 1 diabetes patients. J Exp Med 209(1):51–60

Davies TF, Latif R, Yin X (2012) New genetic insights from autoimmune thyroid disease. J Thyroid Res 2012:623852

Deapen D, Escalante A, Weinrib L, Horwitz D, Bachman B, Roy-Burman P, Walker A, Mack TM (1992) A revised estimate of twin concordance in systemic lupus erythematosus. Arthritis Rheum 35(3):311–318

Degn SE, Jensenius JC, Thiel S (2011) Disease-causing mutations in genes of the complement system. Am J Hum Genet 88(6):689–705

Dennis G Jr, Holweg CT, Kummerfeld SK, Choy DF, Setiadi AF, Hackney JA, Haverty PM, Gilbert H, Lin WY, Diehl L, Fischer S, Song A, Musselman D, Klearman M, Gabay C, Kavanaugh A, Endres J, Fox DA, Martin F, Townsend MJ (2014) Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. Arthritis Res Ther 16(2):R90

Dey-Rao R, Sinha AA (2017) Vitiligo blood transcriptomics provides new insights into disease mechanisms and identifies potential novel therapeutic targets. BMC Genomics 18(1):109

Dotsenko V, Oittinen M, Taavela J, Popp A, Peraaho M, Staff S, Sarin J, Leon F, Isola J, Maki M, Viiri K (2021) Genome-wide transcriptomic analysis of intestinal mucosa in celiac disease patients on a gluten-free diet and postgluten challenge. Cell Mol Gastroenterol Hepatol 11(1):13–32

Dvorak CC, Haddad E, Buckley RH, Cowan MJ, Logan B, Griffith LM, Kohn DB, Pai SY, Notarangelo L, Shearer W, Prockop S, Kapoor N, Heimall J, Chaudhury S, Shyr D, Chandra S, Cuvelier G, Moore T, Shenoy S, Goldman F, Smith AR, Sunkersett G, Vander Lugt M, Caywood E, Quigg T, Torgerson T, Chandrakasan S, Craddock J, Davila Saldana BJ, Gillio A, Shereck E, Aquino V, DeSantes K, Knutsen A, Thakar M, Yu L, Puck JM (2019) The genetic landscape of severe combined immunodeficiency in the United States and Canada in the current era (2010–2018). J Allergy Clin Immunol 143(1):405–407

Eggesbo LM, Risnes LF, Neumann RS, Lundin KEA, Christophersen A, Sollid LM (2020) Single-cell TCR sequencing of gut intraepithelial gammadelta T cells reveals a vast and diverse repertoire in celiac disease. Mucosal Immunol 13(2):313–321

Eriksson D, Royrvik EC, Aranda-Guillen M, Berger AH, Landegren N, Artaza H, Hallgren A, Grytaas MA, Strom S, Bratland E, Botusan IR, Oftedal BE, Breivik L, Vaudel M, Helgeland O, Falorni A, Jorgensen AP, Hulting AL, Svartberg J, Ekwall O, Fougner KJ, Wahlberg J, Nedrebo BG, Dahlqvist P, Norwegian Addison Registry Study Group, Swedish Addison Registry Study Group, Knappskog PM, Wolff ASB, Bensing S, Johansson S, Kampe O, Husebye ES (2021) GWAS for autoimmune Addison's disease identifies multiple risk loci and highlights AIRE in disease susceptibility. Nat Commun 12(1):959

Eugster A, Lindner A, Catani M, Heninger AK, Dahl A, Klemroth S, Kuhn D, Dietz S, Bickle M, Ziegler AG, Bonifacio E (2015) High diversity in the TCR repertoire of GAD65 autoantigen-specific human CD4+ T cells. J Immunol 194(6):2531–2538

Fasano A, Catassi C (2012) Clinical practice. Celiac disease. N Engl J Med 367(25):2419–2426

Ferguson PJ, Blanton SH, Saulsbury FT, McDuffie MJ, Lemahieu V, Gastier JM, Francke U, Borowitz SM, Sutphen JL, Kelly TE (2000) Manifestations and linkage analysis in X-linked autoimmunity-immunodeficiency syndrome. Am J Med Genet 90(5):390–397

Finnish-German AC (1997) An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. Nat Genet 17(4):399–403

Fischer A (2000) Severe combined immunodeficiencies (SCID). Clin Exp Immunol 122(2): 143–149

Fischer A, Lisowska-Grospierre B, Anderson DC, Springer TA (1988) Leukocyte adhesion deficiency: molecular basis and functional consequences. Immunodefic Rev 1(1):39–54

Fisher GH, Rosenberg FJ, Straus SE, Dale JK, Middleton LA, Lin AY, Strober W, Lenardo MJ, Puck JM (1995) Dominant interfering Fas gene mutations impair apoptosis in a human autoimmune lymphoproliferative syndrome. Cell 81(6):935–946

Fuchs YF, Eugster A, Dietz S, Sebelefsky C, Kuhn D, Wilhelm C, Lindner A, Gavrisan A, Knoop J, Dahl A, Ziegler AG, Bonifacio E (2017) CD8(+) T cells specific for the islet autoantigen IGRP are restricted in their T cell receptor chain usage. Sci Rep 7:44661

Galligan CL, Baig E, Bykerk V, Keystone EC, Fish EN (2007) Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: correlates with disease activity. Genes Immun 8(6):480–491

Gangi-Peterson L, Sorscher DH, Reynolds JW, Kepler TB, Mitchell BS (1999) Nucleotide pool imbalance and adenosine deaminase deficiency induce alterations of N-region insertions during V(D)J recombination. J Clin Invest 103(6):833–841

Gardner JM, Devoss JJ, Friedman RS, Wong DJ, Tan YX, Zhou X, Johannes KP, Su MA, Chang HY, Krummel MF, Anderson MS (2008) Deletional tolerance mediated by extrathymic Aire-expressing cells. Science 321(5890):843–847

Gardner JM, Metzger TC, McMahon EJ, Au-Yeung BB, Krawisz AK, Lu W, Price JD, Johannes KP, Satpathy AT, Murphy KM, Tarbell KV, Weiss A, Anderson MS (2013) Extrathymic Aire-expressing cells are a distinct bone marrow-derived population that induce functional inactivation of CD4(+) T cells. Immunity 39(3):560–572

Gavin MA, Torgerson TR, Houston E, DeRoos P, Ho WY, Stray-Pedersen A, Ocheltree EL, Greenberg PD, Ochs HD, Rudensky AY (2006) Single-cell analysis of normal and FOXP3-

mutant human T cells: FOXP3 expression without regulatory T cell development. Proc Natl Acad Sci U S A 103(17):6659–6664

Gavin MA, Rasmussen JP, Fontenot JD, Vasta V, Manganiello VC, Beavo JA, Rudensky AY (2007) Foxp3-dependent programme of regulatory T-cell differentiation. Nature 445(7129):771–775

Giarratana N, Penna G, Adorini L (2007) Animal models of spontaneous autoimmune disease: type 1 diabetes in the nonobese diabetic mouse. Methods Mol Biol 380:285–311

Gioia L, Holt M, Costanzo A, Sharma S, Abe B, Kain L, Nakayama M, Wan X, Su A, Mathews C, Chen YG, Unanue E, Teyton L (2019) Position beta57 of I-A(g7) controls early anti-insulin responses in NOD mice, linking an MHC susceptibility allele to type 1 diabetes onset. Sci Immunol 4(38):eaaw6329

Gottenberg JE, Cagnard N, Lucchesi C, Letourneur F, Mistou S, Lazure T, Jacques S, Ba N, Ittah M, Lepajolec C, Labetoulle M, Ardizzone M, Sibilia J, Fournier C, Chiocchia G, Mariette X (2006) Activation of IFN pathways and plasmacytoid dendritic cell recruitment in target organs of primary Sjogren's syndrome. Proc Natl Acad Sci U S A 103(8):2770–2775

Ha E, Bae SC, Kim K (2020) Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. Ann Rheum Dis 80:558–565

He Y, Xu H, Liang L, Zhan Z, Yang X, Yu X, Ye Y, Sun L (2008) Antiinflammatory effect of Rho kinase blockade via inhibition of NF-kappaB activation in rheumatoid arthritis. Arthritis Rheum 58(11):3366–3376

Heikkila N, Laakso SM, Mannerstrom H, Kekalainen E, Saavalainen P, Jarva H, Arstila TP (2016) Expanded CD4(+) effector/memory T cell subset in APECED produces predominantly interferon gamma. J Clin Immunol 36(6):555–563

Hirschhorn R, Martiniuk F, Rosen FS (1978) Adenosine deaminase activity in normal tissues and tissues from a child with severe combined immunodeficiency and adenosine deaminase deficiency. Clin Immunol Immunopathol 9(3):287–292

Hjelmervik TO, Petersen K, Jonassen I, Jonsson R, Bolstad AI (2005) Gene expression profiling of minor salivary glands clearly distinguishes primary Sjogren's syndrome patients from healthy control subjects. Arthritis Rheum 52(5):1534–1544

Horvath S, Nazmul-Hossain AN, Pollard RP, Kroese FG, Vissink A, Kallenberg CG, Spijkervet FK, Bootsma H, Michie SA, Gorr SU, Peck AB, Cai C, Zhou H, Wong DT (2012) Systems analysis of primary Sjogren's syndrome pathogenesis in salivary glands identifies shared pathways in human and a mouse model. Arthritis Res Ther 14(6):R238

Hui Z, Zhang J, Zheng Y, Yang L, Yu W, An Y, Wei F, Ren X (2021) Single-cell sequencing reveals the transcriptome and TCR characteristics of pTregs and in vitro expanded iTregs. Front Immunol 12:619932

Husebye ES, Anderson MS, Kampe O (2018) Autoimmune polyendocrine syndromes. N Engl J Med 378(12):1132–1141

Inamo J, Suzuki K, Takeshita M, Kassai Y, Takiguchi M, Kurisu R, Okuzono Y, Tasaki S, Yoshimura A, Takeuchi T (2020) Identification of novel genes associated with dysregulation of B cells in patients with primary Sjogren's syndrome. Arthritis Res Ther 22(1):153

Jabri B, Sollid LM (2009) Tissue-mediated control of immunopathology in coeliac disease. Nat Rev Immunol 9(12):858–870

Jacquemin C, Martins C, Lucchese F, Thiolat D, Taieb A, Seneschal J, Boniface K (2020) NKG2D defines a subset of skin effector memory CD8 T cells with proinflammatory functions in Vitiligo. J Invest Dermatol 140(6):1143–1153, e1145

Kaleviste E, Ruhlemann M, Karner J, Haljasmagi L, Tserel L, Org E, Trebusak Podkrajsek K, Battelino T, Bang C, Franke A, Peterson P, Kisand K (2020) IL-22 paucity in APECED is associated with mucosal and microbial alterations in ORAL cavity. Front Immunol 11:838

Kaljas Y, Liu C, Skaldin M, Wu C, Zhou Q, Lu Y, Aksentijevich I, Zavialov AV (2017) Human adenosine deaminases ADA1 and ADA2 bind to different subsets of immune cells. Cell Mol Life Sci 74(3):555–570

Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, Ciclitira PJ, Sollid LM, Partanen J, European Genetics Cluster on Celiac Disease (2003) HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Hum Immunol 64(4):469–477

Katz EL, Harris JE (2021) Translational research in Vitiligo. Front Immunol 12:624517

Kekalainen E, Tuovinen H, Joensuu J, Gylling M, Franssila R, Pontynen N, Talvensaari K, Perheentupa J, Miettinen A, Arstila TP (2007) A defect of regulatory T cells in patients with autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy. J Immunol 178(2):1208–1215

Kendall JL, Springer JM (2020) The many faces of a monogenic autoinflammatory disease: adenosine deaminase 2 deficiency. Curr Rheumatol Rep 22(10):64

Khattri R, Cox T, Yasayko SA, Ramsdell F (2003) An essential role for Scurfin in CD4+CD25+ T regulatory cells. Nat Immunol 4(4):337–342

Kisand K, Link M, Wolff AS, Meager A, Tserel L, Org T, Murumagi A, Uibo R, Willcox N, Trebusak Podkrajsek K, Battelino T, Lobell A, Kampe O, Lima K, Meloni A, Ergun-Longmire B, Maclaren NK, Perheentupa J, Krohn KJ, Scott HS, Husebye ES, Peterson P (2008) Interferon autoantibodies associated with AIRE deficiency decrease the expression of IFN-stimulated genes. Blood 112(7):2657–2666

Kisand K, Boe Wolff AS, Podkrajsek KT, Tserel L, Link M, Kisand KV, Ersvaer E, Perheentupa J, Erichsen MM, Bratanic N, Meloni A, Cetani F, Perniola R, Ergun-Longmire B, Maclaren N, Krohn KJ, Pura M, Schalke B, Strobel P, Leite MI, Battelino T, Husebye ES, Peterson P, Willcox N, Meager A (2010) Chronic mucocutaneous candidiasis in APECED or thymoma patients correlates with autoimmunity to Th17-associated cytokines. J Exp Med 207(2):299–308

Kitagawa Y, Ohkura N, Kidani Y, Vandenbon A, Hirota K, Kawakami R, Yasuda K, Motooka D, Nakamura S, Kondo M, Taniuchi I, Kohwi-Shigematsu T, Sakaguchi S (2017) Guidance of regulatory T cell development by Satb1-dependent super-enhancer establishment. Nat Immunol 18(2):173–183

Knight AK, Cunningham-Rundles C (2006) Inflammatory and autoimmune complications of common variable immune deficiency. Autoimmun Rev 5(2):156–159

Koczan D, Drynda S, Hecker M, Drynda A, Guthke R, Kekow J, Thiesen HJ (2008) Molecular discrimination of responders and nonresponders to anti-TNF alpha therapy in rheumatoid arthritis by etanercept. Arthritis Res Ther 10(3):R50

Koh AS, Miller EL, Buenrostro JD, Moskowitz DM, Wang J, Greenleaf WJ, Chang HY, Crabtree GR (2018) Rapid chromatin repression by Aire provides precise control of immune tolerance. Nat Immunol 19(2):162–172

Kohn DB, Hershfield MS, Puck JM, Aiuti A, Blincoe A, Gaspar HB, Notarangelo LD, Grunebaum E (2019) Consensus approach for the management of severe combined immune deficiency caused by adenosine deaminase deficiency. J Allergy Clin Immunol 143(3):852–863

Koivula TT, Laakso SM, Niemi HJ, Kekalainen E, Laine P, Paulin L, Auvinen P, Arstila TP (2017) Clonal analysis of regulatory T cell defect in patients with autoimmune polyendocrine syndrome type 1 suggests intrathymic impairment. Scand J Immunol 86(4):221–228

Komatsu N, Mariotti-Ferrandiz ME, Wang Y, Malissen B, Waldmann H, Hori S (2009) Heterogeneity of natural Foxp3+ T cells: a committed regulatory T-cell lineage and an uncommitted minor population retaining plasticity. Proc Natl Acad Sci U S A 106(6):1903–1908

Kwon HK, Chen HM, Mathis D, Benoist C (2017) Different molecular complexes that mediate transcriptional induction and repression by FoxP3. Nat Immunol 18(11):1238–1248

Laakso SM, Laurinolli TT, Rossi LH, Lehtoviita A, Sairanen H, Perheentupa J, Kekalainen E, Arstila TP (2010) Regulatory T cell defect in APECED patients is associated with loss of naive FOXP3(+) precursors and impaired activated population. J Autoimmun 35(4):351–357

Laakso SM, Kekalainen E, Rossi LH, Laurinolli TT, Mannerstrom H, Heikkila N, Lehtoviita A, Perheentupa J, Jarva H, Arstila TP (2011) IL-7 dysregulation and loss of CD8+ T cell homeostasis in the monogenic human disease autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy. J Immunol 187(4):2023–2030

Lee BW, Kumar VB, Biswas P, Ko AC, Alameddine RM, Granet DB, Ayyagari R, Kikkawa DO, Korn BS (2018) Transcriptome analysis of orbital adipose tissue in active thyroid eye disease using next generation RNA sequencing technology. Open Ophthalmol J 12:41–52

Leonard MM, Bai Y, Serena G, Nickerson KP, Camhi S, Sturgeon C, Yan S, Fiorentino MR, Katz A, Nath B, Richter J, Sleeman M, Gurer C, Fasano A (2019) RNA sequencing of intestinal mucosa reveals novel pathways functionally linked to celiac disease pathogenesis. PLoS One 14(4):e0215132

Li Q, Wang B, Mu K, Zhang JA (2019) The pathogenesis of thyroid autoimmune diseases: new T lymphocytes – cytokines circuits beyond the Th1-Th2 paradigm. J Cell Physiol 234(3):2204–2216

Lin W, Haribhai D, Relland LM, Truong N, Carlson MR, Williams CB, Chatila TA (2007) Regulatory T cell development in the absence of functional Foxp3. Nat Immunol 8(4):359–368

Liossis SN, Staveri C (2021) What's new in the treatment of systemic Lupus Erythematosus. Front Med (Lausanne) 8:655100

Liston A, Lesage S, Wilson J, Peltonen L, Goodnow CC (2003) Aire regulates negative selection of organ-specific T cells. Nat Immunol 4(4):350–354

Liston A, Gray DH, Lesage S, Fletcher AL, Wilson J, Webster KE, Scott HS, Boyd RL, Peltonen L, Goodnow CC (2004) Gene dosage–limiting role of Aire in thymic expression, clonal deletion, and organ-specific autoimmunity. J Exp Med 200(8):1015–1026

Loberman-Nachum N, Sosnovski K, Di Segni A, Efroni G, Braun T, BenShoshan M, Anafi L, Avivi C, Barshack I, Shouval DS, Denson LA, Amir A, Unger R, Weiss B, Haberman Y (2019) Defining the celiac disease transcriptome using clinical pathology specimens reveals biologic pathways and supports diagnosis. Sci Rep 9(1):16163

Lopez-Herrera G, Tampella G, Pan-Hammarstrom Q, Herholz P, Trujillo-Vargas CM, Phadwal K, Simon AK, Moutschen M, Etzioni A, Mory A, Srugo I, Melamed D, Hultenby K, Liu C, Baronio M, Vitali M, Philippet P, Dideberg V, Aghamohammadi A, Rezaei N, Enright V, Du L, Salzer U, Eibel H, Pfeifer D, Veelken H, Stauss H, Lougaris V, Plebani A, Gertz EM, Schaffer AA, Hammarstrom L, Grimbacher B (2012) Deleterious mutations in LRBA are associated with a syndrome of immune deficiency and autoimmunity. Am J Hum Genet 90(6):986–1001

Lu JM, Chen YC, Ao ZX, Shen J, Zeng CP, Lin X, Peng LP, Zhou R, Wang XF, Peng C, Xiao HM, Zhang K, Deng HW (2019) System network analysis of genomics and transcriptomics data identified type 1 diabetes-associated pathway and genes. Genes Immun 20(6):500–508

Luo J, Liao X, Zhang L, Xu X, Ying S, Yu M, Zhu L, Lin S, Wang X (2020) Transcriptome sequencing reveals potential roles of ICOS in primary Sjogren's syndrome. Front Cell Dev Biol 8:592490

Lvovs D, Favorova OO, Favorov AV (2012) A polygenic approach to the study of polygenic diseases. Acta Nat 4(3):59–71

Ma J, Fu L, Gu H, Chen Z, Zhang J, Zhao S, Zhu X, Liu H, Wu R (2020) Screening for genetic mutations for the early diagnosis of common variable immunodeficiency in children with refractory immune thrombocytopenia: a retrospective data analysis from a Tertiary Children's Center. Front Pediatr 8:595135

Mallone R, Martinuzzi E, Blancou P, Novelli G, Afonso G, Dolz M, Bruno G, Chaillous L, Chatenoud L, Bach JM, van Endert P (2007) CD8+ T-cell responses identify beta-cell autoimmunity in human type 1 diabetes. Diabetes 56(3):613–621

Malmstrom V, Catrina AI, Klareskog L (2017) The immunopathogenesis of seropositive rheumatoid arthritis: from triggering to targeting. Nat Rev Immunol 17(1):60–75

Manuel RC, Pilar BZ, Raphaele S, Hendrika B, Simon JB, Thomas D, Jacques-Eric G, Xavier M, Elke T, Stefano B, Salvatore V, Thomas M, Wan-Fai N, Aike K, Athanasios T, Claudio V, Force ESST (2017) Characterization of systemic disease in primary Sjogren's syndrome: EULAR-SS Task Force recommendations for articular, cutaneous, pulmonary and renal involvements. Rheumatology (Oxford) 56(7):1245

Martinez-Hernandez R, Serrano-Somavilla A, Ramos-Levi A, Sampedro-Nunez M, Lens-Pardo A, Munoz De Nova JL, Trivino JC, Gonzalez MU, Torne L, Casares-Arias J, Martin-Cofreces

NB, Sanchez-Madrid F, Marazuela M (2019) Integrated miRNA and mRNA expression profiling identifies novel targets and pathological mechanisms in autoimmune thyroid diseases. EBioMedicine 50:329–342

McMurchy AN, Gillies J, Gizzi MC, Riba M, Garcia-Manteiga JM, Cittaro D, Lazarevic D, Di Nunzio S, Piras IS, Bulfone A, Roncarolo MG, Stupka E, Bacchetta R, Levings MK (2013) A novel function for FOXP3 in humans: intrinsic regulation of conventional T cells. Blood 121(8):1265–1275

Meredith M, Zemmour D, Mathis D, Benoist C (2015) Aire controls gene expression in the thymic epithelium with ordered stochasticity. Nat Immunol 16(9):942–949

Michels AW, Landry LG, McDaniel KA, Yu L, Campbell-Thompson M, Kwok WW, Jones KL, Gottlieb PA, Kappler JW, Tang Q, Roep BO, Atkinson MA, Mathews CE, Nakayama M (2017) Islet-derived CD4 T cells targeting Proinsulin in human autoimmune diabetes. Diabetes 66(3):722–734

Nagamine K, Peterson P, Scott HS, Kudoh J, Minoshima S, Heino M, Krohn KJ, Lalioti MD, Mullis PE, Antonarakis SE, Kawasaki K, Asakawa S, Ito F, Shimizu N (1997) Positional cloning of the APECED gene. Nat Genet 17(4):393–398

Nature Immunol (2019) The immune cell landscape in kidneys of patients with lupus nephritis PMID: 31209404, PMCID: PMC6726437, 20(7):902–914. https://doi.org/10.1038/s41590-019-0398-x. Epub 2019 Jun 17

Navon Elkan P, Pierce SB, Segel R, Walsh T, Barash J, Padeh S, Zlotogorski A, Berkun Y, Press JJ, Mukamel M, Voth I, Hashkes PJ, Harel L, Hoffer V, Ling E, Yalcinkaya F, Kasapcopur O, Lee MK, Klevit RE, Renbaum P, Weinberg-Shukron A, Sener EF, Schormair B, Zeligson S, Marek-Yagel D, Strom TM, Shohat M, Singer A, Rubinow A, Pras E, Winkelmann J, Tekin M, Anikster Y, King MC, Levy-Lahad E (2014) Mutant adenosine deaminase 2 in a polyarteritis nodosa vasculopathy. N Engl J Med 370(10):921–931

Niemi HJ, Laakso S, Salminen JT, Arstila TP, Tuulasvaara A (2015) A normal T cell receptor beta CDR3 length distribution in patients with APECED. Cell Immunol 295(2):99–104

Oftedal BE, Hellesen A, Erichsen MM, Bratland E, Vardi A, Perheentupa J, Kemp EH, Fiskerstrand T, Viken MK, Weetman AP, Fleishman SJ, Banka S, Newman WG, Sewell WA, Sozaeva LS, Zayats T, Haugarvoll K, Orlova EM, Haavik J, Johansson S, Knappskog PM, Lovas K, Wolff AS, Abramson J, Husebye ES (2015) Dominant mutations in the autoimmune regulator AIRE are associated with common organ-specific autoimmune diseases. Immunity 42(6):1185–1196

Otsubo K, Kanegane H, Kamachi Y, Kobayashi I, Tsuge I, Imaizumi M, Sasahara Y, Hayakawa A, Nozu K, Iijima K, Ito S, Horikawa R, Nagai Y, Takatsu K, Mori H, Ochs HD, Miyawaki T (2011) Identification of FOXP3-negative regulatory T-like (CD4(+)CD25(+)CD127(low)) cells in patients with immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome. Clin Immunol 141(1):111–120

Palomares O, Martin-Fontecha M, Lauener R, Traidl-Hoffmann C, Cavkaytar O, Akdis M, Akdis CA (2014) Regulatory T cells and immune regulation of allergic diseases: roles of IL-10 and TGF-beta. Genes Immun 15(8):511–520

Pandiyan P, Zheng L, Ishihara S, Reed J, Lenardo MJ (2007) CD4+CD25+Foxp3+ regulatory T cells induce cytokine deprivation-mediated apoptosis of effector CD4+ T cells. Nat Immunol 8(12):1353–1362

Park JH, Lee KH, Jeon B, Ochs HD, Lee JS, Gee HY, Seo S, Geum D, Piccirillo CA, Eisenhut M, van der Vliet HJ, Lee JM, Kronbichler A, Ko Y, Shin JI (2020) Immune dysregulation, polyendocrinopathy, enteropathy, X-linked (IPEX) syndrome: a systematic review. Autoimmun Rev 19(6):102526

Passerini L, Olek S, Di Nunzio S, Barzaghi F, Hambleton S, Abinun M, Tommasini A, Vignola S, Cipolli M, Amendola M, Naldini L, Guidi L, Cecconi M, Roncarolo MG, Bacchetta R (2011) Forkhead box protein 3 (FOXP3) mutations lead to increased TH17 cell numbers and regulatory T-cell instability. J Allergy Clin Immunol 128(6):1376–1379 e1371

Pathiraja V, Kuehlich JP, Campbell PD, Krishnamurthy B, Loudovaris T, Coates PT, Brodnicki TC, O'Connell PJ, Kedzierska K, Rodda C, Bergman P, Hill E, Purcell AW, Dudek NL, Thomas

HE, Kay TW, Mannering SI (2015) Proinsulin-specific, HLA-DQ8, and HLA-DQ8-transdimer-restricted CD4+ T cells infiltrate islets in type 1 diabetes. Diabetes 64(1):172–182

Perdriger A, Werner-Leyval S, Rollot-Elamrani K (2003) The genetic basis for systemic lupus erythematosus. Joint Bone Spine 70(2):103–108

Pontynen N, Strengell M, Sillanpaa N, Saharinen J, Ulmanen I, Julkunen I, Peltonen L (2008) Critical immunological pathways are downregulated in APECED patient dendritic cells. J Mol Med (Berl) 86(10):1139–1152

Powell BR, Buist NR, Stenzel P (1982) An X-linked syndrome of diarrhea, polyendocrinopathy, and fatal infection in infancy. J Pediatr 100(5):731–737

Puel A, Doffinger R, Natividad A, Chrabieh M, Barcenas-Morales G, Picard C, Cobat A, Ouachee-Chardin M, Toulon A, Bustamante J, Al-Muhsen S, Al-Owain M, Arkwright PD, Costigan C, McConnell V, Cant AJ, Abinun M, Polak M, Bougneres PF, Kumararatne D, Marodi L, Nahum A, Roifman C, Blanche S, Fischer A, Bodemer C, Abel L, Lilic D, Casanova JL (2010) Autoantibodies against IL-17A, IL-17F, and IL-22 in patients with chronic mucocutaneous candidiasis and autoimmune polyendocrine syndrome type I. J Exp Med 207(2):291–297

Qian L, Shi H, Ding M (2019) Comparative analysis of gene expression profiles in children with type 1 diabetes mellitus. Mol Med Rep 19(5):3989–4000

Rahman A, Isenberg DA (2008) Systemic lupus erythematosus. N Engl J Med 358(9):929–939

Rama M, Duflos C, Melki I, Bessis D, Bonhomme A, Martin H, Doummar D, Valence S, Rodriguez D, Carme E, Genevieve D, Heimdal K, Insalaco A, Franck N, Queyrel-Moranne V, Tieulie N, London J, Uettwiller F, Georgin-Lavialle S, Belot A, Kone-Paut I, Hentgen V, Boursier G, Touitou I, Sarrabay G (2018) A decision tree for the genetic diagnosis of deficiency of adenosine deaminase 2 (DADA2): a French reference centres experience. Eur J Hum Genet 26(7):960–971

Ramsey C, Winqvist O, Puhakka L, Halonen M, Moro A, Kampe O, Eskelin P, Pelto-Huikko M, Peltonen L (2002) Aire deficient mice develop multiple features of APECED phenotype and show altered immune response. Hum Mol Genet 11(4):397–409

Rashighi M, Agarwal P, Richmond JM, Harris TH, Dresser K, Su MW, Zhou Y, Deng A, Hunter CA, Luster AD, Harris JE (2014) CXCL10 is critical for the progression and maintenance of depigmentation in a mouse model of vitiligo. Sci Transl Med 6(223):223ra223

Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X, Alfredsson L, Padyukov L, Klareskog L, Worthington J, Siminovitch KA, Bae SC, Plenge RM, Gregersen PK, de Bakker PI (2012) Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet 44(3):291–296

Remedios KA, Zirak B, Sandoval PM, Lowe MM, Boda D, Henley E, Bhattrai S, Scharschmidt TC, Liao W, Naik HB, Rosenblum MD (2018) The TNFRSF members CD27 and OX40 coordinately limit TH17 differentiation in regulatory T cells. Sci Immunol 3(30):eaau2042

Riding RL, Richmond JM, Fukuda K, Harris JE (2020) Type I interferon signaling limits viral vector priming of CD8(+) T cells during initiation of vitiligo and melanoma immunotherapy. Pigment Cell Melanoma Res 34:683–695

Russell WL, Russell LB, Gower JS (1959) Exceptional inheritance of a sex-linked gene in the mouse explained on the basis that the X/O sex-chromosome constitution is female. Proc Natl Acad Sci U S A 45(4):554–560

Samstein RM, Arvey A, Josefowicz SZ, Peng X, Reynolds A, Sandstrom R, Neph S, Sabo P, Kim JM, Liao W, Li MO, Leslie C, Stamatoyannopoulos JA, Rudensky AY (2012) Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. Cell 151(1):153–166

Sanayama Y, Ikeda K, Saito Y, Kagami S, Yamagata M, Furuta S, Kashiwakuma D, Iwamoto I, Umibe T, Nawata Y, Matsumura R, Sugiyama T, Sueishi M, Hiraguri M, Nonaka K, Ohara O, Nakajima H (2014) Prediction of therapeutic responses to tocilizumab in patients with rheumatoid arthritis: biomarkers identified by analysis of gene expression in peripheral blood mononuclear cells using genome-wide DNA microarray. Arthritis Rheum 66(6):1421–1431

Sansom SN, Shikama-Dorn N, Zhanybekova S, Nusspaumer G, Macaulay IC, Deadman ME, Heger A, Ponting CP, Hollander GA (2014) Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. Genome Res 24(12):1918–1931

Schuetz C, Neven B, Dvorak CC, Leroy S, Ege MJ, Pannicke U, Schwarz K, Schulz AS, Hoenig M, Sparber-Sauer M, Gatz SA, Denzer C, Blanche S, Moshous D, Picard C, Horn BN, de Villartay JP, Cavazzana M, Debatin KM, Friedrich W, Fischer A, Cowan MJ (2014) SCID patients with ARTEMIS vs RAG deficiencies following HCT: increased risk of late toxicity in ARTEMIS-deficient SCID. Blood 123(2):281–289

Schuppan D, Junker Y, Barisani D (2009) Celiac disease: from pathogenesis to novel therapies. Gastroenterology 137(6):1912–1933

Seay HR, Yusko E, Rothweiler SJ, Zhang L, Posgai AL, Campbell-Thompson M, Vignali M, Emerson RO, Kaddis JS, Ko D, Nakayama M, Smith MJ, Cambier JC, Pugliese A, Atkinson MA, Robins HS, Brusko TM (2016) Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. JCI Insight 1(20):e88242

Sellam J, Marion-Thore S, Dumont F, Jacques S, Garchon HJ, Rouanet S, Taoufik Y, Hendel-Chavez H, Sibilia J, Tebib J, Le Loet X, Combe B, Dougados M, Mariette X, Chiocchia G (2014) Use of whole-blood transcriptomic profiling to highlight several pathophysiologic pathways associated with response to rituximab in patients with rheumatoid arthritis: data from a randomized, controlled, open-label trial. Arthritis Rheum 66(8):2015–2025

Sharma S, Pettus J, Gottschalk M, Abe B, Gottlieb P, Teyton L (2019) Single-cell analysis of CD4 T cells in Type 1 diabetes: from mouse to man, how to perform mechanistic studies. Diabetes 68(10):1886–1891

Singh A, Gotherwal V, Junni P, Vijayan V, Tiwari M, Ganju P, Kumar A, Sharma P, Fatima T, Gupta A, Holla A, Kar HK, Khanna S, Thukral L, Malik G, Natarajan K, Gadgil CJ, Lahesmaa R, Natarajan VT, Rani R, Gokhale RS (2017) Mapping architectural and transcriptional alterations in non-lesional and lesional epidermis in vitiligo. Sci Rep 7(1):9860

Smith TJ, Hegedus L (2016) Graves' disease. N Engl J Med 375(16):1552–1565

Sumitomo S, Nagafuchi Y, Tsuchida Y, Tsuchiya H, Ota M, Ishigaki K, Suzuki A, Kochi Y, Fujio K, Yamamoto K (2018) Transcriptome analysis of peripheral blood from patients with rheumatoid arthritis: a systematic review. Inflamm Regen 38:21

Teos LY, Alevizos I (2017) Genetics of Sjogren's syndrome. Clin Immunol 182:41–47

Thomas G, Mancini J, Jourde-Chiche N, Sarlon G, Amoura Z, Harle JR, Jougla E, Chiche L (2014) Mortality associated with systemic lupus erythematosus in France assessed by multiple-cause-of-death analysis. Arthritis Rheum 66(9):2503–2511

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, de la Concha EG, de Almeida RC, Dias KR, van Diemen CC, Dubois PC, Duerr RH, Edkins S, Franke L, Fransen K, Gutierrez J, Heap GA, Hrdlickova B, Hunt S, Izurieta LP, Izzo V, Joosten LA, Langford C, Mazzilli MC, Mein CA, Midah V, Mitrovic M, Mora B, Morelli M, Nutland S, Nunez C, Onengut-Gumuscu S, Pearce K, Platteel M, Polanco I, Potter S, Ribes-Koninckx C, Ricano-Ponce I, Rich SS, Rybak A, Santiago JL, Senapati S, Sood A, Szajewska H, Troncone R, Varade J, Wallace C, Wolters VM, Zhernakova A, D. Spanish Consortium on the Genetics of Coeliac Disease, PreventCD Study Group, Wellcome Trust Case Control Consortium, Thelma BK, Cukrowska B, Urcelay E, Bilbao JR, Mearin ML, Barisani D, Barrett JC, Plagnol V, Deloukas P, Wijmenga C, van Heel DA (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43(12):1193–1201

Tsokos GC (2011) Systemic lupus erythematosus. N Engl J Med 365(22):2110–2121

Tulic MK, Cavazza E, Cheli Y, Jacquel A, Luci C, Cardot-Leccia N, Hadhiri-Bzioueche H, Abbe P, Gesson M, Sormani L, Regazzetti C, Beranger GE, Lereverend C, Pons C, Khemis A, Ballotti R, Bertolotto C, Rocchi S, Passeron T (2019) Innate lymphocyte-induced CXCR3B-mediated melanocyte apoptosis is a potential initiator of T-cell autoreactivity in vitiligo. Nat Commun 10(1):2178

Tuovinen H, Pontynen N, Gylling M, Kekalainen E, Perheentupa J, Miettinen A, Arstila TP (2009) gammadelta T cells develop independently of Aire. Cell Immunol 257(1–2):5–12 van den Wijngaard R, Wankowicz-Kalinska A, Le Poole C, Tigges B, Westerhof W, Das P (2000) Local immune response in skin of generalized vitiligo patients. Destruction of melanocytes is associated with the prominent presence of CLA+ T cells at the perilesional site. Lab Investig 80(8):1299–1309

Van Gool F, Nguyen MLT, Mumbach MR, Satpathy AT, Rosenthal WL, Giacometti S, Le DT, Liu W, Brusko TM, Anderson MS, Rudensky AY, Marson A, Chang HY, Bluestone JA (2019) A mutation in the transcription factor Foxp3 drives T helper 2 effector function in regulatory T cells. Immunity 50(2):362–377, e366

Vihinen M, Mattsson PT, Smith CI (2000) Bruton tyrosine kinase (BTK) in X-linked agammaglobulinemia (XLA). Front Biosci 5:D917–D928

Villa A, Santagata S, Bozzi F, Imberti L, Notarangelo LD (1999) Omenn syndrome: a disorder of Rag1 and Rag2 genes. J Clin Immunol 19(2):87–97

Walker MR, Kasprowicz DJ, Gersuk VH, Benard A, Van Landeghen M, Buckner JH, Ziegler SF (2003) Induction of FoxP3 and acquisition of T regulatory activity by stimulated human CD4+CD25- T cells. J Clin Invest 112(9):1437–1443

Watanabe N, Gao S, Kajigaya S, Diamond C, Alemu L, Ombrello A, Young NS (2019) Analysis of deficiency of adenosine deaminase 2 pathogenesis based on single cell RNA sequencing of monocytes. Blood 134(Supplement_1):409–424

Wildin RS, Ramsdell F, Peake J, Faravelli F, Casanova JL, Buist N, Levy-Lahad E, Mazzella M, Goulet O, Perroni L, Bricarelli FD, Byrne G, McEuen M, Proll S, Appleby M, Brunkow ME (2001) X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. Nat Genet 27(1):18–20

Willemsen M, Melief CJM, Bekkenk MW, Luiten RM (2020) Targeting the PD-1/PD-L1 axis in human Vitiligo. Front Immunol 11:579022

Wilson PC, Wu H, Kirita Y, Uchimura K, Ledru N, Rennke HG, Welling PA, Waikar SS, Humphreys BD (2019) The single-cell transcriptomic landscape of early human diabetic nephropathy. Proc Natl Acad Sci U S A 116(39):19619–19625

Wright HL, Thomas HB, Moots RJ, Edwards SW (2015) Interferon gene expression signature in rheumatoid arthritis neutrophils correlates with a good response to TNFi therapy. Rheumatology (Oxford) 54(1):188–193

Ye H, Zhang J, Wang J, Gao Y, Du Y, Li C, Deng M, Guo J, Li Z (2015) CD4 T-cell transcriptome analysis reveals aberrant regulation of STAT3 and Wnt signaling pathways in rheumatoid arthritis: evidence from a case-control study. Arthritis Res Ther 17:76

Yee CS, Su L, Toescu V, Hickman R, Situnayake D, Bowman S, Farewell V, Gordon C (2015) Birmingham SLE cohort: outcomes of a large inception cohort followed for up to 21 years. Rheumatology (Oxford) 54(5):836–843

Yin X, Sachidanandam R, Morshed S, Latif R, Shi R, Davies TF (2014) mRNA-Seq reveals novel molecular mechanisms and a robust fingerprint in Graves' disease. J Clin Endocrinol Metab 99(10):E2076–E2083

Yu R, Broady R, Huang Y, Wang Y, Yu J, Gao M, Levings M, Wei S, Zhang S, Xu A, Su M, Dutz J, Zhang X, Zhou Y (2012) Transcriptome analysis reveals markers of aberrantly activated innate immunity in vitiligo lesional and non-lesional skin. PLoS One 7(12):e51040

Zakharov PN, Hu H, Wan X, Unanue ER (2020) Single-cell RNA sequencing of murine islets shows high cellular complexity at all stages of autoimmune diabetes. J Exp Med 217(6):e20192362

Zemmour D, Zilionis R, Kiner E, Klein AM, Mathis D, Benoist C (2018) Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. Nat Immunol 19(3):291–301

Zemmour D, Charbonnier LM, Leon J, Six E, Keles S, Delville M, Benamar M, Baris S, Zuber J, Chen K, Neven B, Garcia-Lloret MI, Ruemmele FM, Brugnara C, Cerf-Bensussan N, Rieux-Laucat F, Cavazzana M, Andre I, Chatila TA, Mathis D, Benoist C (2021) Single-cell analysis of FOXP3 deficiencies in humans and mice unmasks intrinsic and extrinsic CD4(+) T cell perturbations. Nat Immunol 22:607–619

Zhang Y, Cai Y, Shi M, Jiang S, Cui S, Wu Y, Gao XH, Chen HD (2016) The prevalence of Vitiligo: a meta-analysis. PLoS One 11(9):e0163806

Zhang Q, Cui T, Chang Y, Zhang W, Li S, He Y, Li B, Liu L, Wang G, Gao T, Li C, Jian Z (2018) HO-1 regulates the function of Treg: association with the immune intolerance in vitiligo. J Cell Mol Med 22(9):4335–4343

Zhang F, Wei K, Slowikowski K, Fonseka CY, Rao DA, Kelly S, Goodman SM, Tabechian D, Hughes LB, Salomon-Escoto K, Watts GFM, Jonsson AH, Rangel-Moreno J, Meednu N, Rozo C, Apruzzese W, Eisenhaure TM, Lieb DJ, Boyle DL, Mandelin AM 2nd, Accelerating Medicines Partnership Rheumatoid Arthritis, Systemic Lupus Erythematosus Consortium, Boyce BF, DiCarlo E, Gravallese EM, Gregersen PK, Moreland L, Firestein GS, Hacohen N, Nusbaum C, Lederer JA, Perlman H, Pitzalis C, Filer A, Holers VM, Bykerk VP, Donlin LT, Anolik JH, Brenner MB, Raychaudhuri S (2019a) Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat Immunol 20(7):928–942

Zhang L, Xu P, Wang X, Zhang Z, Zhao W, Li Z, Yang G, Liu P (2019b) Identification of differentially expressed genes in primary Sjogren's syndrome. J Cell Biochem 120(10):17368–17377

Zhang Z, Xiong D, Wang X, Liu H, Wang T (2021) Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. Nat Methods 18(1):92–99

Zhou Q, Yang D, Ombrello AK, Zavialov AV, Toro C, Zavialov AV, Stone DL, Chae JJ, Rosenzweig SD, Bishop K, Barron KS, Kuehn HS, Hoffmann P, Negro A, Tsai WL, Cowen EW, Pei W, Milner JD, Silvin C, Heller T, Chin DT, Patronas NJ, Barber JS, Lee CC, Wood GM, Ling A, Kelly SJ, Kleiner DE, Mullikin JC, Ganson NJ, Kong HH, Hambleton S, Candotti F, Quezado MM, Calvo KR, Alao H, Barham BK, Jones A, Meschia JF, Worrall BB, Kasner SE, Rich SS, Goldbach-Mansky R, Abinun M, Chalom E, Gotte AC, Punaro M, Pascual V, Verbsky JW, Torgerson TR, Singer NG, Gershon TR, Ozen S, Karadag O, Fleisher TA, Remmers EF, Burgess SM, Moir SL, Gadina M, Sood R, Hershfield MS, Boehm M, Kastner DL, Aksentijevich I (2014) Early-onset stroke and vasculopathy associated with mutations in ADA2. N Engl J Med 370(10):911–920

Ziegler SF (2006) FOXP3: of mice and men. Annu Rev Immunol 24:209–226

Chapter 17
Transcriptome in Human Mycoses

Nalu T. A. Peres, Tamires A. Bitencourt, Gabriela F. Persinoti, Elza A. S. Lang, Antonio Rossi, and Nilce M. Martinez-Rossi

17.1 Introduction

Fungi are eukaryotic microorganisms widely distributed in nature, existing as yeasts, molds, and mushrooms. Fungi are important decomposers of biomass and are useful in baking and wine fermentation. However, fungi can also cause severe, life-threatening infections in humans, animals, and vegetables, resulting in enormous economic losses. Humans are constantly in contact with fungi by inhaling spores in the air and ingesting them as nutritional sources. Human mycoses have increased in incidence due to the high prevalence of immunocompromised patients, becoming a major public-health concern. According to the Global Action Fund for Fungal Infections (GAFFI – https://www.gaffi.org), fungal diseases affect more than 300 million people, leading to the death of approximately 1.6 million people

N. T. A. Peres
Department of Genetics, Ribeirão Preto Medical School, University of São Paulo,
Ribeirão Preto, SP, Brazil

Department of Microbiology, Institute of Biological Sciences, Federal University of Minas
Gerais, Belo Horizonte, MG, Brazil
e-mail: nalu@usp.br

T. A. Bitencourt · E. A. S. Lang · A. Rossi · N. M. Martinez-Rossi (✉)
Department of Genetics, Ribeirão Preto Medical School, University of São Paulo,
Ribeirão Preto, SP, Brazil
e-mail: elzalang@usp.br; anrossi@usp.br; nmmrossi@usp.br

G. F. Persinoti
Department of Genetics, Ribeirão Preto Medical School, University of São Paulo,
Ribeirão Preto, SP, Brazil

Brazilian Biorenewables National Laboratory, Brazilian Center for Research in Energy and
Materials, Campinas, SP, Brazil
e-mail: gabriela.felix@usp.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_17

annually worldwide. Although fungal infections are widespread, they are often overlooked, and in general, public health agencies perform little surveillance of fungal infections (Brown et al. 2012; Rodrigues and Nosanchuk 2020). Fungi cause a wide spectrum of diseases, ranging from asymptomatic infection to disseminated and fatal diseases. Nevertheless, fungal infections are not frequently diagnosed, which impairs the proper epidemiological surveillance of these diseases. The increasing clinical reports of fungal coinfections among hospitalized patients, especially those with respiratory infections such as the recent COVID-19 pandemic caused by the SARS-CoV-2 virus (Zhu et al. 2020; Silva et al. 2021; Alanio et al. 2020), highlight the importance of investments in basic and clinical research exploring fungal mechanisms and pathways. Therefore, more data about the life cycle of pathogenic fungi and the pathogenesis of these infections will aid the development of therapeutic approaches and diagnostic tests. Although research funding for human mycoses remains lower than that for other areas of medical microbiology, the number of publications in the field of medical mycology has increased over the past several decades.

Fungi infect several anatomical sites, resulting in different clinical symptoms. The most prevalent are cutaneous, mucosal, subcutaneous, and pulmonary diseases. These infections can be acquired from trauma to the skin and mucosa, direct or indirect contact with infected humans and animals, contact with contaminated fomites, or inhalation. Airborne fungal infections typically result in pulmonary diseases. Skin and nail infections affect both healthy and immunocompromised individuals, decreasing quality of life by causing discomfort and pruritus. Cutaneous infections are most commonly caused by dermatophytes, a closely related group of keratinophilic molds that infect humans and animals. They are directly or indirectly transmitted between infected organisms and contaminated objects, such as towels and manicure appliances (Peres et al. 2010a). *Candida* species can also cause skin and nail infections, but they more commonly cause oropharyngeal (thrush) and vulvovaginal candidiasis. Many *Candida* species are harmless and are commensal microorganisms. However, immune system impairment favors their pathogenicity, and they can cause opportunistic infections. *C. albicans* is part of the normal microbiota of mucous membranes of the respiratory, gastrointestinal, and female genital tracts. Changes in the host's immunological status and microbiota enable its invasive behavior, leading to tissue damage and dissemination through the bloodstream to other organs (d'Enfert et al. 2020). More recently, the emerging pathogen *Candida auris* poses a threat causing nosocomial infections and hospital outbreaks, increasing global concern owing to its high resistance to antifungal agents and disinfectant chemical compounds (Du et al. 2020).

Deep fungal infections are mainly caused by *Aspergillus fumigatus*, *Cryptococcus neoformans*, *Cryptococcus gattii*, *Coccidioides immitis*, *Paracoccidioides brasiliensis*, *Histoplasma capsulatum,* and *Blastomyces dermatitidis*. Some fungal diseases are endemic, such as blastomycosis (*B. dermatitidis*) and histoplasmosis (*H. capsulatum*), which are mainly found in the United States, and paracoccidioidomycosis (*P. brasiliensis*), which is primarily found in Latin America. Others are cosmopolitan and are encountered worldwide (Brown et al. 2012). Fungal spores

Table 17.1 Main fungal pathogens and their associated diseases in humans

Fungi	Main species	Disease
Aspergillus	*A. fumigauts*	Pulmonary infections (invasive aspergilosis and aspergilloma) Allergy
Blastomyces	*B. dermatitis*	Skin lesions Pulmonary infections
Candida	*C. albicans* *C. glabrata* *C. parapsilosis* *C. auris*	Cutaneous infections – skin and nails oropharyngeal candidiasis (thrush) Vulvovaginal candidiasis Nosocomial infections, fungemia
Coccidioides	*C. immitis* *C. posadasii*	Pulmonary infections
Cryptococcus	*C. neoformans* *C. gattii*	Meningitis Meningoencephalitis Pulmonary infections (pneumonia)
Dermatophytes	*Trichophyton rubrum* *Trichophyton mentagrophytes* *Microsporum canis*	Cutaneous infections – skin, nail, and hair (tinea or ringworms)
Histoplasma	*H. capsulatum*	Pulmonary infections
Malassezia	*M. furfur*	Cutaneous infections – skin (pityriasis versicolor)
Paracoccidioides	*P. brasiliensis* *P. lutzii*	Pulmonary and systemic infections
Penicillium	*P. marneffeii*	Pulmonary infections
Pneumocystis	*P. jirovecii*	Pneumonia
Sporothrix	*S. brasiliensis* *S. schenkii*	Subcutaneous infections

are present in the environment and can be inhaled; upon reaching the lungs, they adhere to the parenchyma and initiate the infectious process. From the lungs, they can enter the bloodstream and disseminate to other organs, mainly the liver and spleen. *Cryptococcus* spp. may also migrate into the central nervous system, causing meningitidis and meningoencephalitis. Table 17.1 summarizes the major human fungal pathogens and their associated diseases. Overall, treatment of clinical mycoses can be a very long and expensive process that is often associated with uncomfortable side effects that lead to treatment interruption (Martinez-Rossi et al. 2018).

Some fungal species are found as filamentous or yeast forms, while others are dimorphic (i.e., found in both forms). In dimorphic fungi, the yeast form represents the parasitic phase, and the hyphae form represents the saprophytic phase. The filament (or hyphae) is a tubular multicellular structure, and cells may be divided into compartments by the formation of a septum. Yeasts are round, single cells that reproduce by budding and some species can form pseudohyphae, a chain of interconnected yeast cells. Fungi can undergo sexual or asexual reproduction, producing spores that can be inhaled or enter the body at sites of tissue damage. Once the spores reach their appropriate niche, they develop into hyphae that invade the tissue in search of nutrients.

The genomes of several fungal pathogens have been sequenced, enabling the design and analysis of microarrays, high-throughput RNA sequencing (RNA-seq), and RT-qPCR (Reverse Transcription – quantitative Polymerase Chain Reaction). Fungal transcriptomics has been used to analyze gene expression and regulation in response to antifungal exposure, environmental changes, and interaction with the host during infection. The transcriptional profile may help elucidate several aspects of fungal biology, including signaling pathways that enable fungal survival and help predict molecular targets for the development of novel antifungal drugs (Peres et al. 2010b; Cairns et al. 2010). Transcriptional and proteomic analyses have been used to identify connections among signaling and metabolic pathways that govern fungal development, morphogenesis, antifungal resistance, and pathogenicity as well as the host's immune response. Recent advances in molecular biology methods and bioinformatics tools have enabled the study of the whole transcriptome, providing meaningful insights into the functionality of the genome, revealing interaction networks and molecular components of cells and tissues involved in physiological and pathological processes. Transcriptome allows the analysis of all transcript species, including mRNAs, which encode proteins, as well as noncoding RNAs (ncRNAs) and small RNAs (sRNAs), which regulate gene expression and maintain cellular homeostasis. Furthermore, transcriptional profiling by RNA-seq is useful to determine gene structures at transcription initiation sites, 5′- and 3′-ends, and introns as well as splicing patterns (Stark et al. 2019).

This chapter will discuss recent advances in fungal transcriptomics arising from microarray and RNA-seq analyses. The contribution of these findings to the understanding of fungal biology and fungal diseases will be highlighted. The intrinsic relationship between the outcome of fungal infections and the immunological status of the host stresses the need to evaluate the host immune response to fungi. Furthermore, knowledge of the genes expressed in response to stressful environmental conditions and the gene networks that regulate the transcriptome during a fungal infection will help elucidate the pathogenesis of fungal infections and identify possible molecular targets for the development of novel therapeutic agents. Such information will aid both the treatment and prevention of fungal infections.

17.2 Host Immune Response to Fungal Infections

The host's immunological status is the primary determinant of the severity of fungal infections, which can range from asymptomatic to severe and disseminated. Immunocompromised patients often suffer from severe, disseminated, and fatal fungal infections. Host–pathogen interactions are complex and involve several molecules on the surface of both host and fungal cells. Understanding the infective process requires molecular knowledge of the pathogen strategies for infecting the tissue, as well as the host responses aimed at eliminating the pathogen and maintaining cellular integrity. The development of experimental models has improved the study of infectious diseases, and most of these models utilize immunosuppressed mice because most fungal species cause opportunistic infections. However, for some pathogens, such as anthropophilic dermatophytes, these models are not suitable and *ex vivo* and *in vitro* assays have been performed, providing insights into the pathogenic process and immune response triggered by the fungus.

Fungal diseases can result from poor immune responses or from exacerbated activation of the immune system, such as the inflammatory response. Therefore, the interplay of the innate and adaptive immune mechanisms and their appropriate activation are crucial for successful pathogen clearance and cellular homeostasis. The innate immune response is comprised of the epithelial barrier, mucosa, and phagocytes (i.e., neutrophils, macrophages, and dendritic cells [DCs]), which play essential roles in preventing the entry of pathogenic microorganisms and rapidly killing these pathogens, as well as activating the adaptive immune response. Complement and other molecules, such as antimicrobial peptides (AMPs) and mannose-binding lectin, are also important host defense mechanisms. Pattern-recognition receptors (PRRs) on the surface of host cells interact with pathogen-associated molecular patterns (PAMPs), such as α- and β-glucans, mannans, lipopolysaccharides, and phospholipomannan in the fungal cell wall. The molecular interaction between PRRs and PAMPs triggers intracellular signaling pathways that initiate early inflammatory and non-specific responses in the host and upregulates virulence factors in the pathogen that enhance survival. PRRs include the toll-like receptors (TLRs) TLR2, TLR4, and TLR9, complement receptor 3, mannose receptor, Fcγ receptor, Dectin-1 and 2, Galectin, Macrophage-Inducible C-type Lectin (mincle), and Dendritic Cells - Specific Intercellular adhesion molecule Grabbing Non-integrin (DC-SIGN) (Romani 2011). Pathogen recognition by macrophages leads to their differentiation into the classic (M1) and alternatively (M2) activated macrophages. While M1 macrophages have microbicidal and pro-inflammatory properties, playing a role in fungal clearance, M2 macrophages have anti-inflammatory activities implicated in fungal persistence (Pathakumari et al. 2020). Furthermore, besides fungal PAMPs, recent studies on fungal extracellular vesicles have demonstrated their role in modulating the host immune response (Bitencourt et al. 2018; Bielska and May 2019).

In general, a Th1 response is correlated with protective immunity against fungi, and is characterized by the production of interferon gamma (IFN-γ), among other cytokines, leading to cell-mediated immunity. Antigen-presenting cells (APCs), such as macrophages and DCs, initiate the Th1 response once their PRRs engage with fungal PAMPS, which leads to cellular activation and elicits effector properties. Th1 cells are essential for optimal activation of phagocytes at the site of infection through the production of signature cytokines. Moreover, Th17 cells support Th1 cellular responses, playing an important role in promoting neutrophil recruitment (Bedoya et al. 2013; Pathakumari et al. 2020; Romani 2011). Regulatory T cells (Treg) control tissue damage by reducing the inflammatory response. However, this also causes immune suppression, thereby allowing fungal persistence. Th2 response, characterized by the production of IL-4, IL-5, IL-10, TGF-β, and IL-13, is correlated with an increased fungal burden. However, IL-4 may play a protective role in the early stages of fungal infections, and the balance between Th1 and Th2 is crucial for the outcome of the infection (Pathakumari et al. 2020). Whole-genome transcriptional analyses have identified specific transcriptional profiles of host cells in response to various fungal species. Different cell types respond to fungal stimuli by activating distinct intracellular signaling pathways downstream of different PRRs. This mechanism confers plasticity to immune cells, such as DCs and macrophages, which shapes T-cell responses during fungal infections. The distinct signaling pathways in phagocytes influence the balance between innate and adaptive immune responses and the balance between CD4^{+} T cells and Treg, establishing the outcome of the infection (Romani 2011).

H. capsulatum is a dimorphic fungus that causes respiratory infections and disseminated disease in immunocompromised hosts. The *H. capsulatum* hyphae produce spores (conidia) in the environment, which can be inhaled by humans. Inside the host, the conidia undergo morphological changes to form yeast cells. Once inhaled, the conidia are captured by phagocytic cells, such as macrophages, and trigger the host immune response. However, yeast cells use alveolar macrophages as vehicles to spread to different organs, such as the liver, spleen, lymph nodes, and bone marrow (Mittal et al. 2019). Microarray analysis revealed that in response to conidia, macrophages specifically upregulated type I IFN-induced genes, including IFN-β and a classic type 1 IFN signature, in addition to general inflammatory genes. This effect was dependent on interferon regulatory factor 3 (IRF3) and independent of the TLR signaling pathway. IFNAR1 (type I IFN receptor) knockout mice showed a decreased fungal burden in the lungs and spleen after intranasal infection with conidia and yeast cells, compared to wild-type mice. Therefore, IFNAR1 signaling might contribute to disease and fungal burden rather than conferring protection, through the modulation of cytokines, apoptosis of infected macrophages, or specific aspects of the adaptive immune response to *H. capsulatum* (Inglis et al. 2010). However, in the pathogenic yeast *C. neoformans*, which causes severe meningoencephalitis in immunocompromised patients, type 1 IFN signaling directs cytokine responses toward a protective type 1 pattern during murine cryptococcosis. IFNAR1 and IFN-β knockout infected mice displayed higher fungal burdens in the lungs and brain and decreased survival, compared to wild-type mice (Biondo et al. 2008). Likewise, *C. albicans* induced the expression of type 1 IFN genes and proteins in DCs but not in macrophages. IFNAR1 and IFN-β knockout mice also displayed a lower survival rate and increased fungal burden in the kidneys, showing that type 1 IFN response plays a protective role against *C. albicans* (Biondo et al. 2011).

The major virulence factor of *C. neoformans* is its polysaccharide capsule, which interferes with recognition by immune cells. Microarray analysis showed that a nonencapsulated strain induced the expression of genes involved in DC maturation, chemokines, and cytokines, characterizing an immunostimulatory response. Among the proteins encoded by these upregulated genes were CD86, CD83, the transcription factor Relb, ICAM1, major histocompatibility complex class II (MHC-II)-related genes (H2-D1, H2-Q7, and H2-Q8), and the pro-inflammatory cytokines IL-12, TNF-α, and IL-1. Several chemokines were also upregulated in DCs stimulated with the nonencapsulated strain, including CCL3, CCL4, CCL7, CCL12, CXCL10, CCL22, and the chemokine receptor CCR7, which contributes to the accumulation of inflammatory cells at the site of infection. In contrast, an encapsulated strain caused downregulation or no change in the expression of these genes, indicating that the capsule prevented the activation of immune response-related genes. Among the proteins encoded by the genes downregulated by the encapsulated strain were E74-like factor 1 (Elf1) and sequestosome 1 (Sqstm1), which regulate the expression of cytokines genes and the induction of NF-κB signaling, respectively (Lupo et al. 2008). Additionally, *C. neoformans* profoundly alters the expression profiles of both polarized macrophages M1 and M2 to a naive phenotype (Subramani et al. 2020) under *in vitro* conditions. RNA-seq comparing pulmonary infection by *C. neoformans* in mice and monkeys revealed species-specific responses. The expression of genes coding for IL-1a and IL-1b and those involved in iron acquisition, transport, and storage was upregulated in the lungs of monkeys. However, the expression of calcium homeostasis-related genes was repressed in mice but remained unchanged in monkeys. Genes related to the TLR, TNF, IL-17 pathways, and copper homeostasis were upregulated in both models; the insulin signaling pathway was also modulated in response to *C. neoformans* infection (Li et al. 2019).

P. brasiliensis causes pulmonary and systemic infections. Microarray analyses of murine macrophages and DCs after phagocytosis of *P. brasiliensis* identified differential expression of genes encoding inflammatory cytokines, chemokines, signal-transduction proteins, and apoptosis-related proteins (Silva et al. 2008). Among the genes upregulated in macrophages were the pro-inflammatory chemokines CCL21, CCL22, and CXCL1. CXCL1 and CCL22 recruit neutrophils and monocytes, respectively, while CCL21 mediates the homing of lymphocytes to secondary lymphoid organs. Upregulation of the gene encoding NF-κB might account for the upregulation of pro-inflammatory chemokines and cytokines (e.g., TNF-α) that increase the cytotoxic activity of macrophages (Silva et al. 2008). Also, TNF-α deficient mice were unable to control *P. brasiliensis* infection, given the increased fungal burden and the absence of a well-formed granuloma (Silva et al. 2008; Souto et al. 2000). After exposure to *P. brasiliensis*, macrophages highly expressed apoptotic genes, including caspases 2, 3, and 8, which may represent a mechanism of eliminating the fungus without damaging host tissues. On the other hand, the fungus induced the expression of matrix metalloproteases genes, which may have facilitated fungal invasion, given their role in tissue remodeling (Silva et al. 2008).

Expression of the gene encoding IL-12 was downregulated in macrophages interacting with *P. brasiliensis*; however, it was upregulated in DCs interacting with *P. brasiliensis* (Tavares et al. 2012) and *C. neoformans* (Lupo et al. 2008). IL-12 is associated with resistance to paracoccidioidomycosis and cryptococcosis by inducing IFN-γ production and Th1 protective responses. IL-12p40 knockout mice displayed decreased survival, higher fungal burden, and decreased production of IFN-γ (Livonesi et al. 2008). In addition to IL-12, DCs exposed to *P. brasiliensis* expressed genes encoding other pro-inflammatory cytokine and chemokine genes, such as TNF-α, CCL22, CCL27, CXCL10, and NF-κB, concomitantly with the downregulation of the NF-κB inhibitor Nκ-RF encoding gene. Both macrophages and DCs expressed CCL22 in response to *P. brasiliensis*, which might have increased the microbicidal activity of macrophages by stimulating a respiratory burst and the release of lysosomal enzymes. The chemokines expressed by macrophages and DCs in response to *P. brasiliensis* mediate the accumulation of leukocytes at the site of infection in order to control fungal invasion (Silva et al. 2008; Tavares et al. 2012). There is a significant difference between the transcriptome profile of DCs derived from susceptible and resistant mice infected *in vitro* with *P. brasiliensis*. This observation highlights that a high activation of the inflammatory responses and downregulation of autophagy, lysosome, and apoptosis are involved in the disease. A low activation of these pathways is related to infection resistance and a proper immune response (de-Souza-Silva et al. 2020).

The dynamics of the molecular response triggered by *C. albicans* in human monocytes identified a pattern of gene expression related to recruitment, activation, and viability of phagocytes, as well as the enhancement of chemotaxis and inflammation (Kim et al. 2005). Increased expression of genes encoding TNF-α, IL-6, and IL-1α was correlated with neutrophil infiltration at the site of infection. There was an upregulation of genes encoding the chemokines CCL3, CCL4, CCL20, CCL18, CXCL1, CXCL-3, and IL-8, which are involved in the activation and recruitment of phagocytes and lymphocytes, as well as genes encoding the chemokine receptors CCR1, CCR5, CCR7, and CXCR5. In the early stages of infection, monocytes overexpress genes encoding various pro-inflammatory cytokines, chemokines, and chemokine receptors as well as COX2, IL-23, which are important for inflammation, and heat-shock proteins, which are implicated in the induction of inflammatory cytokines and chemokines. Thus, these changes in gene expression allow cellular recruitment and activation. Along with the pro-inflammatory response, increased expression of genes encoding anti-apoptotic molecules (XIAP and BCL2A1) may have protected the monocytes from cellular damage and death. The gene encoding the transferrin receptor (CD71) was upregulated, suggesting that iron deprivation might be a defense mechanism against infection. Indeed, iron is essential to the virulence of several pathogens (Johns et al. 2021). Further, RNA-seq of mononuclear cells challenged with *C. albicans* and *C. auris* showed unique and species-specific transcriptional signatures. *C. auris* induced the expression of type I and II IFNs, IFN-related genes, IL-1RA, IL-10, IL-9, and IL-27, thus triggering a stronger host response than *C. albicans*. This may be attributed to differences in the cell wall mannoproteins, leading to different phagocytic indices and clinical outcomes between these two species (Bruno et al. 2020).

Neutrophils display a potent set of hydrolytic enzymes, antimicrobial peptides, and oxidative species within their intracellular granules, having an immediate and pronounced effect on *C. albicans* (Fradin et al. 2005). Granulocyte-like cells phagocytose and kill *C. albicans*, prevent hyphal growth, and undergo apoptosis after pathogen exposure. During this process, granulocytes upregulate inflammatory genes and downregulate anti-*Candida*-response genes, depending on the size of the inoculum. Among the upregulated genes were inflammatory mediators, including IL-1β, TNF-α, COX2, and the chemokine CCL3. On the other hand, genes encoding myeloperoxidase, which causes hyphal damage, and defensins, such as human neutrophil protein 1 (HPN1), were downregulated. These changes may represent mechanisms by which *C. albicans* survives the early stages of infection (Mullick et al.

2004). In another study, a microarray of immune-related genes was used to evaluate the early response of PMN cells to *C. albicans* hyphal cells, UV-killed and live yeasts. In PMNs, the transcriptional profiles induced by live yeasts and hyphae were more similar to one another than to that induced by dead yeasts. This suggested that fungal viability had a more significant effect on PMN gene expression than cellular morphotype. The presence of *C. albicans* did not affect the expression of genes encoding granule proteins. Nevertheless, *C. albicans* induced the upregulation of pro-inflammatory genes and cell-to-cell signaling (leukemia inhibitory factor [LIF]), signal transduction proteins, cell stimulatory factors, vascular endothelial growth factor, and PMN-recruitment chemokines (CCL3 and CXCL2). Importantly, these gene expression changes were irrespective of fungal cell type or viability. Furthermore, the few genes that were downregulated in response to *C. albicans* were involved in the regulation of cell signaling and growth (Fradin et al. 2007). In addition, exposure to viable *Candida* cells upregulated genes encoding stress-response proteins, including heat shock proteins (HSPA8, HSPCA, HSPCB, and HSPH1). This demonstrated a direct effect of live cells on PMN cells and monocytes (Kim et al. 2005). Interestingly, these genes also regulate CXC-type chemokines, indicating that this antimicrobial response amplifies the overall immune response by recruiting additional cells to the infection site. Overall, this transcriptional profile suggested that PMNs contribute to the immunological response to *C. albicans* by expressing genes involved in cellular communication, which may recruit more PMNs or other immune cells.

Systemic candidiasis is characterized by *C. albicans* entering the bloodstream, disseminating throughout the body, and causing microabscesses. In the blood vessels, the fungus adheres and invades endothelial cells (ECs); thus, the ECs have the potential to influence the host response to vascular invasion. Microarray transcriptional analysis of ECs in response to *C. albicans* identified the upregulation of genes involved in chemotaxis, angiogenesis, cell death, proliferation, intra- and intercellular signaling, immune response, and inflammation (Muller et al. 2007; Barker et al. 2008; Lim et al. 2011). *C. albicans* induces several genes that are targets of the pro-inflammatory transcription factor NF-κB, and chemokines, including IL-8, CXCL1, CXCL2, CXCL3, CXCL5, and CXCL6, indicating that ECs help to recruit neutrophils and monocytes to the infection site (Muller et al. 2007). The overexpression of genes involved in stress and wound healing, such as IL-1, calgranulin C, E-selectin, and prostaglandin-endoperoxide synthase 2, correlated with the endothelial damage caused by *C. albicans*. The ECs also upregulated antiapoptotic genes, suggesting that ECs respond to *C. albicans* by undergoing cellular proliferation (Barker et al. 2008). However, another transcriptional profile revealed that apoptotic genes were upregulated in ECs infected with a high density of *C. albicans*. In general, human umbilical vein ECs infected with high densities of *C. albicans* displayed a stronger and broader transcriptional response than cells infected with low densities, which may be related to the number of cells or even to secreted molecules involved in quorum sensing. The authors hypothesized that in microenvironments with a high density of yeast cells, such as microabscesses, the fungus triggers apoptosis, which disrupts the endothelial barrier and permits fungal dissemination to different organs and tissues (Lim et al. 2011).

Given the complex and dynamic nature of host–pathogen interactions, techniques that measure both the host and pathogen responses are crucial for characterizing their interaction. Dual transcriptomics is used to identify molecular patterns of the pathogen and the host simultaneously, providing insights into the dynamics of the infectious process (Westermann et al. 2017). Dual transcriptomics by RNA-seq was performed during DC phagocytosis of *C. albicans*, and gene interactions were predicted using systems biology. RNA-seq identified 545 *C. albicans* and 240 DCs genes differentially expressed, clustered by their expression kinetics over the duration of the interaction, and selected genes were used to infer gene interactions. After experimentally validating one of these gene interactions, the authors proposed a model in which PTX3, an opsonin secreted by DCs that facilitates phagocytosis through dectin-1, binds to the *C. albicans* cell wall, leading to its remodeling, which is mediated by the transcription factor Hap3 during the invasion of innate immune cells. Remodeling of the fungal cell wall compromises the ability of immune cells to recognize fungi, thus attenuating the immune response (Tierney et al. 2012).

Microarray-based dual transcriptomics was performed on *A. fumigatus* interacting with bronchial epithelial cells, revealing expression patterns indicating the activation of the host's innate immune response (Oosthuizen et al. 2011). *A. fumigatus* is a major cause of pulmonary fungal infections, including invasive aspergillosis, aspergilloma, and allergy. During infection, environmental conidia enter the airways through inhalation. There, they germinate into hyphae and penetrate the lung parenchyma. Upon invasion, the fungus can disseminate to other organs and tissues. In response to *A. fumigatus* conidia, bronchial cells upregulated genes involved in innate immunity, chemokine activity, and inflammation. Among the overexpressed genes were those encoding the chemokines CCL3 and CCL5, which recruit leukocytes to the site of infection, matrix metalloproteinases (MMP1 and MMP3), and glutathione transferase (MGST1), which protects against oxidative damage. By comparing the expression profiles of two different cell lines, the authors identified only 17 genes in common. This demonstrates the variability in gene expression between a cell line and primary cells resulting from exposure to the same fungus (Gomez et al. 2011; Oosthuizen et al. 2011). The commonly expressed genes mainly encoded chemokines and regulators of the innate immune response. IL-6, a potent pro-inflammatory cytokine, was highly expressed in response to *A. fumigatus* conidia, consistent with earlier findings that IL-6-deficient mice were susceptible to invasive pulmonary aspergillosis and had impaired protective Th1 responses (Oosthuizen et al. 2011). In addition, genes involved in nucleosome organization and chromatin assembly were overexpressed. Genes involved in mitosis and cell cycle progression were downregulated, suggesting decreased proliferation and cell cycle arrest during infection with *A. fumigatus* (Gomez et al. 2011).

Moreover, in response to *A. fumigatus* human monocytes presented a coordinated expression of genes involved in fungal death and invasion (Cortez et al. 2006). Among the highly expressed genes were pro-inflammatory genes, such as IL-1β, CCL3, CCL4, IL-8, PTX3, and SOD2, and regulators of inflammation, such as

IL-10, COX2, and HSP40. Moreover, several anti-inflammatory genes were downregulated, such as CD14, which is involved in phagocytosis, and CCL5, which is a Th1 chemokine. This differential regulation of pro- and anti-inflammatory genes likely balanced the innate immune response. Furthermore, the coordinated expression of genes involved in oxidative response may have both eliminated the fungus and protected the cell. This was supported by the high expression of superoxide dismutase (SOD2) and dual phosphatase (DUSP1) and downregulation of catalase (CAT), glutathione peroxidase 3 (GPX3), and peroxiredoxin 5 (PRDX5) (Cortez et al. 2006). Recently, a triple RNA-seq analysis of DCs coinfected with *A. fumigatus* and cytomegalovirus (CMV) revealed unique transcription profiles of the host in response to each pathogen alone or during coinfection along with distinct profiles of both pathogens during infection and coinfection. The gene expression profile of DCs showed a different pattern in response to each pathogen. *A. fumigatus* induced Th17 expression, whereas CMV infection led to a Th1 response. However, coinfection led to downregulation of the expression of these genes, with each pathogen attenuating the effect of the other on the molecular signature of DCs and thereby interfering with the host response (Seelbinder et al. 2020).

Dermatophytes are highly specialized fungi that use keratin as a nutrient source, and thus infect keratinized structures, such as skin, hair, and nails. Upon infecting the skin, dermatophytes first encounter keratinocytes, which represent an important barrier against pathogens and help mediate the immune response (Burstein et al. 2020). The zoophilic dermatophyte *Arthroderma benhamiae*, and the anthropophilic species *Trichophyton tonsurans* induced different cytokine expression profiles in keratinocytes, correlating with the inflammatory response. In infected keratinocytes, the zoophilic species induced the upregulation of pro-inflammatory genes and the concomitant secretion of cytokines IL-1β, IL-6, IL-6R, and IL-17 and chemokines IL-8 and CCL2. This effect may promote the infiltration of inflammatory cells in the skin during infection, and the upregulation of IL-6, IL-6R, and granulocyte-colony stimulating factor (G-CSF) may lead to tissue remodeling and wound healing. On the other hand, the anthropophilic species induced limited cytokine expression and release, including exotoxin 2, IL-8, and IL-16. This was likely responsible for the poor inflammatory response observed in *T. tonsurans* skin infection (Shiraki et al. 2006). Moreover, mice infected with *A. benhamiae* displayed an infiltration of PMNs, macrophages, and DCs in the skin as well as increased levels of TGF-β, IL-1β, IL-6, and IL-22 mRNA in skin biopsies (Cambier et al. 2014). This pro-inflammatory profile was also observed in keratinocytes challenged with the zoophilic dermatophyte *Microsporum gypseum*, with the upregulation of the expression of IL-6, IL-8, IL-1β, TNF-α, and c-Jun and enrichment of the NF-kB, TNF, and MAPK signaling pathways. However, the gene network of keratinocytes response to *Trichophyton rubrum* was based on metabolic pathways such as steroid, fatty acid, and isoprenoid biosynthesis (Deng et al. 2020). *In vitro* infection of keratinocytes with *T. rubrum* induced the expression of genes coding for the AMPs RNase 7, beta-defensin 3, and the natural resistance-associated macrophage protein 1, in addition to that of cytokine-related genes (Firat et al. 2014); Petrucelli et al. 2018). The expression of genes involved in epidermal cell differentiation, such as caspase 14 and laminin subunit gamma 2, and cell migration, such as metalloproteinase 9, was upregulated, whereas that of genes involved in skin barrier maintenance, such as keratin 1 and filaggrin, was downregulated. This may also account for the tissue damage and antifungal response during *T. rubrum* infection (Petrucelli et al. 2018). Further, the inflammatory response during infection may also be regulated by microRNAs, owing to their upregulation in macrophages challenged with heat-inactivated *T. rubrum* conidia (Gonzalez Segura et al. 2020).

In summary, fungal pathogens induce several changes in the host's target cells and innate immune cells. Studying the transcriptome of fungal–host interactions has elucidated the molecular patterns associated with protection from or progression of fungal infections. In general, fungi induce the upregulation of genes encoding cytokines, chemokines, and other pro-inflammatory molecules in host cells, which recruit inflammatory cells to the site of infection. Host cells exhibit different expression profiles in response to different fungal pathogens, which may account for the differences in outcomes of these infections. Moreover, some studies have identified molecular strategies by which fungi evade the host's immune system as well as host defense mechanisms that favor fungal survival. Transcriptomic analyses have generated hypotheses that can be further validated by reverse genetic approaches to better characterize the immune components that contribute to the outcome of fungal infections.

17.3 Metabolic Adaptation of Fungi During Infection

Fungal pathogens adapt to the host's microenvironment during infection, a process that requires dynamic responses to constantly changing conditions (Brown et al. 2014). In particular, nutrient availability can be limited in host niches, especially inside phagocytes. Host cues and tissue nutrients substantially affect the outcome of the infection by triggering the activation of different fungal signaling pathways that govern germination, cell wall remodeling, and morphological cell type switch, as well as of those regulating the production of enzymes involved in transcription regulation and metabolic adjustments to improve growth and host invasion and dissemination (Johns et al. 2021). Thus, fungi undergo metabolic adaptations to control, for example, glycolysis, gluconeogenesis, glyoxylate cycle, and proteolysis. This allows them to utilize diverse substrates as nutrient sources, evade the toxic conditions triggered by the immune response, and maintain their virulence despite changes in the physiological ambient (Brock 2009). To avoid immune recognition, fungal cells monitor host cues through plasma membrane receptors. Subsequently, they mask cell wall components such as β-glucans and melanin, either by the encapsulation or formation of titan cells, as shown for *C. neoformans*, or by the activation of the transcription factors Crz1 and Ace2 that govern the cell wall remodeling in *C. albicans*, aiding fungal colonization and reducing neutrophil recruitment (Ballou et al. 2016). Moreover, it is well known that phagocytes produce reactive oxygen and nitrogen species (ROS and RNS), which induce oxidative and nitrosative stress as an attempt to kill pathogens (Brown et al. 2009). Reactive species can alter or inactivate proteins, lipid membranes, and DNA. Pathogens can survive this toxic environment by producing protective enzymes, such as flavohemoglobin and S-nitrosoglutathione (GSNO) reductase, which confer resistance to nitrosative stress (de Jesus-Berrios et al. 2003), and superoxide dismutases, catalases, and peroxidases, which counteract oxidative stress. Nonenzymatic defenses include metabolites, such as melanin, mannitol, and trehalose (Missall et al. 2004). The ability of pathogens to sense and appropriately respond to environmental pH is essential for their survival in different host niches. In pathogenic fungi, the PACC/RIM signaling pathway has been implicated in survival, growth, virulence, and dissemination in different host niches (Cornet and Gaillardin 2014; Martinez-Rossi et al. 2017). The pH affects enzymatic activities; the alkaline pH of human tissues influences nutrient uptake because the solubility of essential elements, such as iron and zinc, are pH dependent (Amich et al. 2010). Iron is a critical micronutrient in both the host and pathogen, as it is required for several metabolic processes, including respiration and DNA replication. In the form of heme and iron-sulfur compounds, iron is an essential cofactor in various cellular enzymes, oxygen carriers, and electron-transfer systems. Iron homeostasis plays a key role in host–pathogen interactions. Similarly, zinc is essential for pathogenic fungi because it is a constituent of many transcription factors and acts as a cofactor for enzymes involved in cell signaling. For instance, host tissues can restrict free iron and zinc availability to prevent infection. Accordingly, fungal pathogens have adapted strategies for iron uptake, including the production of metalloreductases, ferroxidases, and siderophores (Silva et al. 2011) and uptake of zinc through the production of zincophore such as Pra1 and its ortholog Aspf2 (Amich et al. 2010; Citiulo et al. 2012) to survive in iron and zinc-deficient niches.

Several pathways are crucial for fungal pathogens to survive in various host microenvironments during infection. *In vivo*, *ex vivo*, and *in vitro* infection models have identified fungal pathogens' transcriptional profiles during infection and interaction with host cells. These studies have helped to elucidate the pathogenesis of superficial, deep, and bloodstream fungal infections. In this sense, an *in vitro* study used microarray to assess the transcriptional profile of *C. albicans* during interaction with human blood. There was an upregulation of genes involved in stress response, such as *SSA4* (a member of the *HSP70* gene family), and anti-oxidative response, such as those encoding Cu/Zn superoxide dismutase (*SOD1*), catalase (*CAT1*), and thioredoxin reductase (*TRR1*). There was a simultaneous upregulation of genes encoding the glycolytic enzymes phosphofructokinase (*PFK2*), phosphoglycerate kinase (*PGK1*), and enolase (*ENO1*), as well as those encoding the glyoxylate cycle enzymes isocitrate lyase (*ICL1*), malate synthase (*MLS1*), and acetyl-coenzyme-A-synthetase (*ACS1*). Genes involved in fermentation, such as those encoding alcohol dehydrogenases (*ADH1* and *ADH2*), were also upregulated. Importantly, *C. albicans* isolated from infected mice exhibited a similar transcription profile, thus validating some of the *in vitro* results. Moreover, these data suggested that *C. albicans* use alternative carbon sources during blood infection and dissemination (Fradin et al. 2003).

A subsequent study investigated the utilization of the glyoxylate cycle and glycolysis by *C. albicans* interacting with different blood fractions, including erythrocytes, PMNs (mainly neutrophils), PMN-depleted blood (consisting of lymphocytes and monocytes), and plasma. *C. albicans* cells were physiologically active and displayed rapid hyphal growth while interacting with plasma, erythrocytes, and PMN-depleted blood. On the other hand, growth of *C. albicans* cells was arrested when interacting with PMNs, and only 40% of the cells interacting with whole blood produced hyphae. *C. albicans* upregulated glyoxylate cycle genes when interacting with PMNs, but not when interacted with plasma. During interaction with plasma and PMN-depleted blood, *C. albicans* upregulated genes related to glycolysis. Global cluster analysis was used to compare the transcriptional profile of *C. albicans* interacting with whole blood and blood fractions. During interaction with whole blood, the upregulation of genes related to glycolysis and the glyoxylate cycle resulted from mixed populations of fungal cells that were internalized by phagocytes, which triggers a nutrient limitation response, and not internalized (Fradin et al. 2005). Indeed, starvation inside the phagosome activated the glyoxylate cycle, which allowed nutrient uptake and survival.

Besides, during the interaction of *C. albicans* with neutrophils, the activation of nitrogen- and carbohydrate-starvation responses was observed, as indicated by the upregulation of genes encoding ammonium permeases (*MEP2* and *MEP3*), vacuolar proteases (*PRB1*, *PRB2*, and *APR1*), carboxypeptidases (*PRC1* and *PRC2*), glyoxylate cycle enzymes (*MLS1*, *ICL1*, and *ACS1*), amino acid transporters, and proteins involved in amino acid metabolism. Moreover, *C. albicans* internalized by murine macrophages *in vitro* displayed growth arrest and downregulation of the expression of genes associated with translation machinery and glycolysis. In contrast, there was an upregulation of genes encoding enzymes involved in the gluconeogenesis (phosphoenolpyruvate carboxykinase and fructose-1,6-bisphosphatase), glyoxylate cycle (isocitrate lyase and malate synthase), tricarboxylic acid cycle (aconitase, citrate synthase, and malate dehydrogenase), and β-oxidation of fatty acids, as well as several transporters. Accordingly, *C. albicans* deficient in the gene encoding isocitrate lyase was less virulent than the wild-type strain in murine infection (Lorenz and Fink 2001). The interaction with host cells also triggered the upregulation of oxidative stress response genes such as superoxide dismutases (*SOD1* and *SOD5*) and catalase (*CAT1*) (Fradin et al. 2005), flavohemoglobin, cytochrome *c* peroxidase, peroxidases, reductases, stress response (heat shock protein HSP78), metal homeostasis, and DNA repair (Lorenz et al. 2004). In accordance with this data, a previous work evaluated the transcriptional profile of *C. albicans* using biopsies of infected oral mucosa from 11 HIV-positive patients showed changes that reflected fungal protective responses toward nitrosative stress, innate defense of epithelial cells against microbes, adaptation to the neutral-alkaline pH of the oral mucosa, and the use of alternative carbon sources at the site of infection. From this evidence in association with literature support, glyoxylate genes have been considered an important virulence factor. Indeed, Δ*icl1* was impaired to damage RHE, suggesting the importance of the glyoxylate cycle in oral candidiasis (Wachtler et al. 2011). Moreover, epithelial escape and dissemination (*EED1*), a unique species-specific *C. albicans* gene, is involved in hyphal elongation during infection (Zakikhany et al. 2007). Time-course microarray analysis of the wild type and Δ*eed1* strains interacting with RHE showed the downregulation of seven genes throughout infection, including the hyphae-associated genes *ECE1* and *HYR1* and those encoding proteins involved in polarized growth, such as CDC42, RDI1, MYO2, CDC11, CYB2, MOB1, and MLC1.

Another study showed the coordinated host and fungal transcriptional response during macrophage infection with *C. albicans* (Munoz et al. 2019). In this study, sorted cells were analyzed with respect to different infection stages, and a single-cell approach was employed to track different trajectories in the course of infection. Important changes in metabolic pathways were evidenced by the expression of genes modulated in phagocytosed *C. albicans*. Alternative pathways were activated at early infection stages to cope with the macrophage environment and favor fungal survival with limited glucose availability in the phagosome, including genes belonging to the glyoxylate cycle and beta-oxidation. In contrast, the expression of genes related to chaperones, transcription factors that regulate translation, and peptide synthesis was downregulated. A shift in gene expression occurred at a later infection time. The central carbon pathway was reactivated with the expression of genes involved in morphological changes, such as cell wall assembly and filamentation. The genes related to phagocytosis and innate immune response activation were enriched in macrophages, including those associated with IL-6, IL-8, and NF-κB signaling pathways. Moreover, the production of nitric oxide, reactive oxygen species (ROS), and pattern recognition receptors was also induced. Notably, there was a shift during the time of the infection, with significant repression of the immune response. Interestingly, morphological changes in *C. albicans* and induction of filamentation were followed by repression of the immune response. Beyond that, the single-cell analysis also provided new insights into infection outcomes. It demonstrated that bimodality in gene expression was observed in about 15% of differentially expressed genes (DEGs) mainly involved in pathogen recognition and pro-inflammatory pathways within 2 h and 4 h of interaction. In *C. albicans,* about 23% of DEGs presented bimodality in gene expression, evidenced mainly in genes related to metabolism and virulence, and hyphae transition. Finally, this study evaluated the alternative splice (AS) as exon skipping in these sorted cells, highlighting the occurrence of two isoforms for a gene that encodes a dectin potentially involved in the Th17 response. From that perspective, both cells trigger mechanisms to promote stochastic diversification to favor the phenotype and infection outcome during the interaction.

In order to investigate expression changes in *C. albicans* during systemic infection, transcriptional profiling was performed *in vivo* on mice infected as well as pig livers inoculated *ex vivo*. The upregulated genes encoded enzymes involved in glycolysis, such as phosphofructokinase (*PFK2*) and pyruvate dehydrogenase subunits (*PDA1* and *PDX1*), as well as those involved in acetyl-CoA biosynthesis and the tricarboxylic acid cycle (*KGD1* and *KGD2*). This gene expression modulation reflected the availability of carbohydrates and the utilization of glycolysis and respiration for energy production. However, the upregulation of *PCK1*, which encodes phosphoenolpyruvate carboxykinase, a key enzyme in gluconeogenesis, suggested that alternative carbon sources were also used. Other upregulated genes included *SAP2*, *SAP4, SAP5, and SAP6*, which encode the hyphae-associated aspartic proteases. Indeed, Sap2 is the major protease that enables the utilization of proteins as nitrogen sources. The upregulation of alkaline pH responsive gene (*PHR1*) suggested adaptation to an alkaline environment. Similarly, upregulation of genes encoding stress-response proteins, including heat shock proteins and molecular chaperones (*HSP78*, *HSP90*, *DDR48*, *HSP104*, *HSP12*, and *SSA4*), suggested that the heat shock response was triggered during the course of infection. However, genes related to oxidative, osmotic, and nitrosative stress were not upregulated. On the other hand, genes related to iron, copper, zinc, and phosphate transport (*FTR1, CTR1, ZRT1, PHO84, PHO89*) were upregulated during liver infection, suggesting limited iron and phosphate in this environment. Also, among genes identified during the comparison of transcriptional profile of an invasive *C. albicans* strain with a noninvasive strain was DFG16, which encodes a membrane sensor in the RIM101 pathway that is crucial for pH-dependent hyphal formation, pH sensing, invasion at physiological pH, and systemic infection (Thewes et al. 2007; Martinez-Rossi et al. 2012; Rossi et al. 2013). Moreover, in rabbits, the infected kidneys with *C. albicans* exhibited an upregulation of genes related to alternative pathways of carbon assimilation, such as β-oxidation of fatty acids, the glyoxylate cycle (*MLS1* and *ACS1*), and the tricarboxylic acid cycle (*CIT1*, *ACO1*, and *SDH12*), suggesting limited carbohydrate supply in the kidneys (Walker et al. 2009). Although genes involved in β-oxidation of fatty acids are upregulated in several infection models, fatty acid degradation is not essential for the virulence of *C. albicans*. Nevertheless, disruption of genes involved in the glyoxylate cycle or gluconeogenesis significantly attenuated its virulence in mice (Ramirez and Lorenz 2007; Barelle et al. 2006).

C. albicans colonizes medical devices, such as intravascular catheters, by forming biofilms. Biofilms are comprised of heterogeneous microbial communities and form on biotic or abiotic surfaces embedded in an extracellular polymeric matrix. Such biofilms are associated with persistent infections and resistance to antifungal drugs and mechanical treatments (Cavalheiro and Teixeira 2018). *C. albicans* forms a biofilm in four steps. First, yeast cells attach to and colonize a surface; second, yeast cells form a basal layer that anchors the biofilm; third, hyphae grow and produce pseudohyphae and extracellular matrix; finally, the yeast cells disperse. In order to characterize biofilm formation in *C. albicans*, the transcriptional regulatory network was analyzed in mutants that are unable to form biofilms. A combination of whole-genome chromatin immunoprecipitation microarray (ChIP-chip) and genome-wide transcriptional profiling identified six master regulators that control biofilm formation in *C. albicans*: BCR1, TEC1, EFG1, NDT80, ROB1, and BRG1. Each regulator controlled the other five, and most of the target genes were controlled by more than one master regulator. However, a recent comprehensive analysis in the circuit of biofilm formation demonstrated that the biofilm/hyphae regulatory network shows a more profound variation in accordance to genotype from each isolate, which was partly attributed to the occurrence of single nucleotide polymorphisms in cis-regulatory elements of BRG1 that influences its control by

BCR1 (Huang et al. 2019). Moreover, the biofilm network targeted approximately 15% of the entire genome (Nobile et al. 2012). The enriched GO terms of EFG1 responsive genes involve biofilm formation and cell surface. It is remarkable that the involvement of carbohydrate metabolism, mainly glycolytic and gluconeogenesis, processes as an important set of EFG1-activated genes (Huang et al. 2019).

In *C. auris,* resistance to different antifungal compounds in association with the capability to form biofilm and rapid dissemination between patients urge for efforts aiming to unveil its physiology. Recent reports have profiled the gene expression of *C. auris* during biofilm formation and exposure to caspofungin (Kean et al. 2018; Zamith-Miranda et al. 2020). In biofilms, the over-enriched GO terms involved translation, siderophore transport, and iron homeostasis. Moreover, in biofilms, the expression of glycosylphosphatidylinositol (GPI)-anchored cell wall genes, potentially involved with adhesion properties, such as *IFF4*, *CSA1*, *PGA26*, and *PGA52*, was upregulated. During the intermediate to mature biofilm formation phase, the expression of some genes encoding efflux pumps, such as *RDC3*, *SNQ2*, *CDR1*, and *YHD3*, was upregulated. Similarly, in mature biofilms, the expression of two adhesin-encoding genes, *HYR3* and *ALS5,* was upregulated (Kean et al. 2018). Another recent study assessed the transcriptional profile of *C. auris* after caspofungin exposure and compared the effect on extracellular vesicles (EVs) secretion. The enriched GO terms involved cell wall biogenesis, cell cycle, oxidative stress response, and protein transport. In addition, morphological topography of yeast cells was affected after caspofungin treatment, and clumps were evidenced, suggesting a transition from yeast to hyphae as a compensatory mechanism to overcome disturbances in the cell wall. Regarding EV production, there was a shift in the content of small RNAs (Zamith-Miranda et al. 2020).

Microarray analyses were used to profile *C. neoformans* transcription profile in response to murine macrophages. *C. neoformans* exhibited a downregulation of genes encoding translational machinery and an upregulation of genes associated with lipid degradation and fatty acid catabolism (lipases and acetyl coenzyme A acetyltransferase), β-oxidation, transport of glucose and other carbohydrates, response to nitrogen starvation, the glyoxylate cycle (*ICL1*), and autophagy (*ATG3* and *ATG9*). Moreover, the upregulation of several genes encoding oxidoreductases, peroxidases, and flavohemoglobin denitrosylase (FHB1), which are important for nitrosative response and virulence, indicated the presence of oxidative and nitrosative stress (de Jesus-Berrios et al. 2003). Also, there was an upregulation of genes related to endocytosis, exocytosis, and synthesis of extracellular polysaccharides and cell wall components. Genes located in the mating-type (MAT) locus and several genes associated with virulence were also upregulated. These included those encoding inositol-phosphorylceramide synthase (*IPC1*), laccases (*LAC1* and *LAC2*), genes involved in capsule formation (*CAP10, CAS31, CAS32, CAS1*, and *CAS2*), and *PKA*, a gene in the Gpa1-cAMP pathway, that is essential for virulence. In particular, the Gpa1-cAMP pathway regulates capsule formation and melanin production. Moreover, calcineurin gene (*CNA1*), which is critical for virulence, was upregulated (Fan et al. 2005).

Transcriptional analyses of *C. neoformans* isolated from cryptococcal pulmonary infection in mice revealed the upregulation of genes encoding malate synthase, phosphoenolpyruvate carboxykinase, aconitase and succinate dehydrogenase as well as those involved in β-oxidation of fatty acids. Genes encoding glyoxylate cycle enzymes were strongly upregulated as well as genes involved in glycolysis (e.g., fructose 1,6-biphosphate, aldolase, hexokinase, and phosphofructokinase). In addition, there was an upregulation of genes encoding transporters for monosaccharides, iron, copper, acetate, trehalose, and phosphate, enzymes involved in the production of acetyl-CoA (e.g., acetylCoA synthetase [*ACS1*]), pyruvate decarboxylase, and aldehyde dehydrogenase. Moreover, the upregulation of several stress-response genes, including flavohemoglobin denitrosylase, superoxide dismutase, *HSP12*, *HSP90*, and other virulence factors were evidenced. Deletion of the *acs1* gene resulted in attenuated virulence and impaired growth on media containing acetate as a carbon source. Moreover, *ACS1* is regulated by serine/threonine protein kinase 1 (SNF1), which mediates glucose sensing, utilization of alternative carbon sources, and stress response. Deletion of the *SNF1* gene also reduced growth on acetate medium, decreased melanin production, and caused loss of virulence in murine model (Hu et al. 2008). Although *C. neoformans* upregulated glyoxylate cycle genes during infection, *ICL1* and *MLS1* were not essential for establishing infection (Rude et al. 2002; Idnurm et al. 2007). On the other hand, deficits in β-oxidation pathways compromised the virulence of *C. neoformans* (Kretschmer et al. 2012).

The gene expression profile was compared among seven isolates of *C. neoformans* var. *grubii,* including the VNI and VNB lineages, comprising four clinical and three environmental isolates grown in five different *in vitro* and *in vivo* conditions (Yu et al. 2020). The conditions corresponded to synthetic media, such as YPD and a restrictive low iron medium supplement with an inductor of ROS, infection models like macrophage-like murine cells infected with yeasts, cerebrospinal fluid (CSF) obtained from intracisternal yeast-infected rabbit, and pigeon media guano, which is a niche environment in terms of nitrogen composition for VNI isolates. Genes that favor fungal virulence and survival within the host were modulated. Typically, genes involved in oxidative stress response, acquisition and reduction of iron, capsule production, glycosylation pathway, ATP-binding cassette transporters, *APP1, CXD3,* and *SRX1* were regulated. *APP1* encodes a secreted anti-phagocytic protein. Although APP1 deletion does not impair the growth, capsule, or melanin production of *C. neoformans*, it increased the dissemination of *C. neoformans* in hosts with compromised immune response, and the AP1 administration inhibited phagocytosis of fungal cells (Luberto et al. 2003). *CXD3* encodes a carboxypeptidase D that seems to participate in nitrogen metabolism and capsule formation (Frazzitta et al. 2013). The sulfiredoxin SRX1 is a virulence factor of *C. neoformans,* which also plays a protective role by counteracting the stress caused by peroxide (Upadhya et al. 2013). Notably, *in vivo* conditions highlighted the expression of metabolic and stress adaptive mechanisms. The functional enrichment of DEGs demonstrated genes involved in amino acid biosynthesis and nitrogen metabolism, cell cycle, DNA repair, stress responses, inositol phosphate metabolism, and inositol lipid modifications. As the brain microenvironment has limited glucose availability, it is conceivable that *C. neoformans* utilize inositol as a carbon source. This study evidenced the upregulation of the expression of virulence-associated genes. Almost half of such genes consisted of capsule production genes. Moreover, DEGs are involved in sodium efflux transport (*ENA1* and *NHA1*), oligopeptide transport, and quorum sensing (*OPT1)*. The expression of stress-responsive genes (*SRE1*and *SREBP*) was upregulated. Similarly, the expression of genes involved in signaling pathways that respond to thermal stress and pH, including Bck1-Mkk2-Mpk1 and the pH-response transcription factor, *RIM101* was upregulated. Beyond that, the targeted genes of RIM101 include *CDA1* and *KRE6*, which are responsible for regulating the levels of chitosan and β-glucan synthesis, respectively, in the cells (O'Meara et al. 2013).

In a murine model of pulmonary aspergillosis, *A. fumigatus* exhibited downregulation of genes related to ribosomal biogenesis and protein biosynthesis and upregulation of approximately 150 genes related to siderophore biosynthesis and transport, including ferric-chelate reductases, amino acid permeases, GABA and proline permeases, maltose permeases and transporters, and extracellular proteases. Elastinolytic metalloprotease, an aorsin-like serine protease, and dipeptidylpeptidases are antigenic virulence factors that are important for nitrogen uptake, and several genes encoding antioxidant enzymes, including a Mn-superoxide dismutase and the bifunctional catalase-peroxidase CAT2 (Oosthuizen et al. 2011). The initiation of infection was likely associated with aminoacid catabolism, as indicated by the induction of the enzyme methylcitrate synthase, which detoxifies propionyl-CoA intermediates, which is a toxic product generated from the degradation of the host aminoacids methione, valine, and isoleucine (McDonagh et al. 2008). Moreover, an *A. fumigatus* strain deficient in methylcitrate synthase displayed attenuated virulence (Ibrahim-Granet et al. 2008). Besides, a current study assessed the transcriptional profile of *A. fumigatus* in a model of invasive pulmonary infection by NanoString nCounter. Among the evaluated genes, the expression of 125 genes was upregulated whereas that of 85 was downregulated, representing genes potentially involved in the response to environmental cues, with an extensive list of transcription factors. Upregulated genes were involved in iron acquisition (*fre2*, *hapX*, *sidA*, *sidD, mirB*, and *sit1),* zinc uptake *(zrfC*, *zrfA*, *aspf2*, and *zafA),* and nitrogen uptake *(nrtB* and *area)*. Simultaneously, the expression of *sreA* was downregulated, whose codified product represses iron uptake and siderophore synthesis. A prominent induction of the expression of genes involved in secondary metabolism, such as *gliG*, *gliP, gliZ* (belonging to gliotoxin biosynthesis pathway), and *mtfA*, which acts in both gliotoxin and extracellular proteases synthesis, was verified. Notably, this study highlighted the role of *rlmA,* which is involved in the ability of fungus to proliferate in the lung, *ace1*, which controls gene clusters related to multiple secondary metabolites, and mycotoxin, which is paramount for full virulence (Liu et al. 2021). Conceivably, the invading hyphae in the lungs trigger neutrophil recruitment, and as a consequence, the fungus activates a stress-responsive mechanism, including the induction of *sebA, mkk2,* and *sho1*. Furthermore, while interacting with human neutrophils, *A. fumigatus* conidia upregulated genes encoding proteins involved in peroxisome biogenesis, β-oxidation of fatty acids (acyl-CoA dehydrogenase and enoyl-CoA hydratase), acetate metabolism (acetyl-coenzyme A synthetase), the tricarboxylic acid cycle (aconitate, succinate dehydrogenase, and malate dehydrogenase), and the glyoxylate cycle (isocitrate lyase) (Sugui et al. 2008). There was a strong upregulation of the gene encoding formate dehydrogenase, which detoxifies formate, an indirect product of the glyoxylate cycle. Albeit the involvement of ROS release by phagocytes in killing *A. fumigatus,* a triple *SOD1/SOD2/SOD3* mutant and the parental strain were similarly virulent in experimental murine aspergillosis in immunocompromised animals (Lambou et al. 2010).

Transcriptional profiling was performed on the dermatophyte *A. benhamiae* during an *in vivo* skin infection in guinea pigs. During acute infection, *A. benhamiae* upregulated genes encoding key enzymes of the glyoxylate cycle (MLS and ICL), formate dehydrogenase, monosaccharide transporter, oxidoreductase, opsin-related protein, and several proteases (Staib et al. 2010). The most highly upregulated gene was *SUB6* that encodes subtilisin 6, a protease previously characterized as the major allergen in another dermatophyte, *T. rubrum*. Sub6 has been shown to bind human IgE antibodies (Woodfolk and Platts-Mills 1998). The second most highly upregulated gene was that encoding an opsin-related protein with an unknown function. Genes encoding proteases, such as subtilisins SUB1, SUB2, SUB6, and SUB7, the neutral protease NpII-1, and serine carboxypeptidase ScpC were also upregulated during infection (Staib et al. 2010). Proteases are the most commonly studied virulence factors of dermatophytes, and their function in generating short peptides and amino acid breakdown products allows them to infect the skin and nails (Monod 2008). Genes encoding SUB3, SUB5, and metalloprotease 4 (MEP4) were also upregulated in *T. rubrum* grown in keratin as the sole carbon source (Maranhão et al. 2007). Moreover, a *PACC/RIM101*-mutant strain of *T. rubrum* displayed decreased keratinolytic activity and impaired growth on the human nail *in* vitro, suggesting a role for RIM101 in the pathogenicity of *T. rubrum* (Ferreira-Nozawa et al. 2006; Silveira et al. 2010; Martinez-Rossi et al. 2012). In addition, the cross-talk of PacC with different pathways for the maintenance of cellular homeostasis has been demonstrated. In *T. interdigitale* the regulation of *egr2* that encodes a C2H2 transcription factor involved in ion homeostasis, and *P-type ATPase* gene, putatively involved in the extrusion of Na+ and K+, is influenced by *pacC* background (da Silva et al. 2020). Moreover, the ortholog of PacC, Pac3 in *N. crassa* is involved in a myriad of processes. The responsive genes involve those that encode catalase 1 and catalase 3, cell wall protein PhiA, C6 transcription factor, calcium-transporting ATPase 3, cyclin, and ornithine N5 oxygenase (Martins et al. 2020a).

Indeed, many aspects of dermatophyte physiology were understood based on studies performed using protein sources like keratin and elastin to mimic the dermatophyte superficial and deep infection, respectively, or even human molecules such as nail fragments and skin explants (Peres et al. 2016). In this sense, a study that assessed the transcriptional profile of *T. rubrum* mycelium grown in keratin or elastin through oligonucleotide microarray displayed the modulation of a large set of proteases, with a significant upregulation of *mep4* and *lap1* genes as well as genes encoding heat shock proteins, including Hsp 70 like-protein, Hsp 88-like protein, and Hsp 90 like-protein (Bitencourt et al. 2019a). Besides, this study showed the equal importance of lipases and keratinases for dermatophyte infection during interaction with elastin. It also revealed the modulation of a large set of genes involved in carbon and nitrogen metabolism. In this context, another study evaluating the transcriptional profile of *T. rubrum* conidia during growth in keratin and elastin revealed the modulation of genes involved in conidia dormancy. Moreover, the expression of protease genes including *lap1, lap2, sub1, sub3, sub6*, and *mep4*, as well as genes belonging to the respiratory chain and tricarboxylic cycle was primarily induced during growth in keratin. In contrast, the expression of approximately 40 genes involved in metabolic processes was downregulated in both protein sources, including genes related to nitrogen and fatty acid metabolism. This study unveiled adaptative mechanisms related to conidia survival and germination and characterized a putative adhesin potentially involved in the initial phases of dermatophyte infection (Bitencourt et al. 2016). In addition, a recent study evaluated the transcriptional profile of *T. rubrum* time-course mycelial growth in minimal medium supplement with glucose or keratin, revealing that keratin growth led to the repression of genes related to glycolysis, nitrogen catabolism, and TCA cycle, and induction of glyoxylate genes, such as *icl* (Martins et al. 2020b). This study also showed that keratin degradation is followed by an accumulation of ammonium, and as a consequence, mechanisms related to glutamine and urea metabolism are activated for ammonium utilization and extrusion.

During interaction with human keratinocytes, *A. benhamiae* upregulated the *hypA* gene, which encodes a hydrophobin (Burmester et al. 2011) that influences the organism's recognition by the immune system (Heddergott et al. 2012). Deletion of *hypA gene* increased the susceptibility of *A. benhamiae* to human neutrophils and DCs. Compared to wild type, the Δ*hypA* mutant strain activated cellular immune defenses and increased the release of IL-6, IL-8, IL-10, and TNF-α to a higher degree. Moreover, conidia of the mutant strain were more easily killed by neutrophils (Heddergott et al. 2012). Indeed, surface expression of hydrophobin prevents *A. fumigatus* recognition by neutrophils (Aimanianda et al. 2009). Furthermore, in *T. rubrum*, the *hypA* gene is regulated by the transcription factor StuA, which belongs to the APSES family of transcription factors. In comparison to the wild type, the Δ*stuA* mutant strain displayed a significant reduction in *hypA* transcript levels during growth in keratin, and as a consequence, altered other mechanisms that ultimately influence germination, stress response, and fungal mechanosensing (Lang et al. 2020). Conceivably, the interaction with host cells and the dampening of host recognition might also be affected. Within this context, LysM-domain proteins influence fungal infection and immune response by masking fungal chitin recognition by host cells and controlling fungal growth. These proteins have garnered interest in dermatophytes due to high copies in the genome and diversification in domain organization, suggesting LysM family evolution in these pathogenic fungi (Martinez et al. 2012; Persinoti et al. 2014). Moreover, a recent work characterized LysM-domain proteins in *T. rubrum* and showed the transcriptional profile of 14 LysM-encoding genes during the growth of *T. rubrum* in host molecules. In this study, two genes (TERG 03756 and TERG 05625) demonstrated marked changes in transcription levels during *T. rubrum* growth in keratin, displaying a signal peptide, hydrophobic region, and two LysM domains without glycosylation sites (Lopes et al. 2019).

Recently, the transcriptional profiles of HaCat keratinocyte cell line and *T. rubrum* during fungus–host interaction have been identified simultaneously. In this respect, dual RNA seq data showed the induction of the expression of glyoxylate cycle genes (malate synthase and isocitrate lyase), *erg6,* and a carboxylic acid transporter gene that probably enhances the assimilation of nutrients (Petrucelli et al. 2018). Furthermore, deletion of *hacA* impaired the hyphal development during interaction with HaCat and altered immune responses, with an increase in TNF-α secretion and a decrease in IL-8 levels (Bitencourt et al. 2020).

Transcriptome data from various human fungal pathogens have identified global responses and survival strategies during interaction with host cells and substrates. Moreover, a deeper understanding of the core in transcriptional responses obtained by analyzing the dynamic and coordinate behavior of fungi and hosts during the infection and tackling aspects of niche association and adaptation is critical for identifying vulnerabilities. Accordingly, some pathways have been implicated in mycotic diseases, fungi can proliferate and survive within the host by employing sophisticated mechanisms to quickly modulate gene expression and adapt to changes in the environment. Genes that are upregulated during the infective process or interaction with host cells are potentially important for virulence (Table 17.2), and the functional characterization of mutant strains has been performed for some of them. Thus, genome-wide transcriptional analyses combined with genetic approaches have provided significant insight into fungal responses, adaptive processes, virulence, and pathogenesis.

17.4 Transcriptome of Drug Response and Resistance

Microorganisms respond to sublethal doses of chemical and physical agents by synthesizing various specific proteins and low molecular weight compounds that act to promote defenses or tolerance (Fachin et al. 2001). Fungi use numerous signal transduction pathways to sense environmental stress and respond appropriately by differentially expressing cell-stress genes (Martinez-Rossi et al. 2018). Thus, analyses of transcriptional changes in response to cytotoxic drugs have elucidated the mechanisms by which fungi adapt to physiological stress and the mechanisms of drug action (Table 17.3).

Although there are several commercially available antifungal drugs, the number of cellular targets is limited. Some antifungal drugs target ergosterol, a sterol analogous to cholesterol that is the main component of the fungal cell membrane and has diverse functions, including maintaining membrane stability, integrity, and permeability. Polyenes, a class of antifungal drugs including amphotericin B (AMB) and nystatin, bind to ergosterol and form pores in the membrane, which leads to the

Table 17.2 Putative fungal proteins associated with host interaction and pathogenesis

Protein description	Gene expression modulation and functional analysis	References
Isocitrate lyase (glyoxylate cycle enzyme)	Upregulated in *C. albicans, C. neoformans*, *A. fumigatus, A. benhamiae,* and *T. rubrum.* Gene inactivation attenuates virulence in *C. albicans* but not in *C. neoformans* and *A. fumigatus.*	Fradin et al. (2003, 2005), Lorenz et al. (2004), Zakikhany et al. (2007), Fan et al. (2005), Chen et al. (2014), Sugui et al. (2008), Staib et al. (2010), Rude et al. (2002), Schobel et al. (2007), Lorenz and Fink (2001), Wachtler et al. (2011), and Martins et al. (2020b)
Malate synthase (glyoxylate cycle enzyme)	Upregulated in *C. albicans, C. neoformans*, *A. fumigatus,* and *A. benhamiae.* Gene inactivation does not attenuate virulence in *C. neoformans.*	Fradin et al. (2003, 2005), Lorenz et al. (2004), Zakikhany et al. (2007), Walker et al. (2009), Hu et al. (2008), McDonagh et al. (2008), Staib et al. (2010), Idnurm et al. (2007), and Cairns et al. (2010)
Acetyl-coenzyme-A-synthetase (glyoxylate cycle enzyme)	Upregulated in *C. albicans, C. neoformans,* and *A. fumigatus.* Gene inactivation attenuates virulence in *C. neoformans.*	Fradin et al. (2003, 2005), Walker et al. (2009), Sugui et al. (2008), McDonagh et al. (2008), Cairns et al. (2010), Hu et al. (2008), Thewes et al. (2007), and Lorenz et al. (2004)
Aconitase (tricarboxylic acid cycle enzyme)	Upregulated in *C. albicans, C. neoformans,* and *A. fumigatus.*	Lorenz et al. (2004), Walker et al. (2009), Hu et al. (2008), and Sugui et al. (2008)
Malate dehydrogenase (tricarboxylic acid cycle enzyme)	Upregulated in *C. albicans, C. neoformans,* and *A. fumigatus.*	Lorenz et al. (2004), Hu et al. (2008), Cairns et al. (2010), and Sugui et al. (2008)
Phosphofrucktokinase (glycolysis enzyme)	Upregulated in *C. albicans* and *C. neoformans*	Fradin et al. (2003), Thewes et al. (2007), and Hu et al. (2008)
Enolase (glycolysis enzyme)	Upregulated in *C. albicans* and *C. neoformans*	Fradin et al. (2003), Thewes et al. (2007), and Hu et al. (2008)
Phosphoenolpyruvate carboxykinase (gluconeogenesis enzyme)	Upregulated in *C. albicans* and *C. neoformans*. Gene inactivation attenuates virulence in *C. albicans*.	Zakikhany et al. (2007), Lorenz et al. (2004), Thewes et al. (2007), Hu et al. (2008), and Barelle et al. (2006)
Flavohemoglobin denitrosylases (RNS detoxification)	Upregulated in *C. albicans* and *C. neoformans*. Gene inactivation attenuates virulence in *C. albicans* and *C. neoformans*.	Hu et al. (2008), Lorenz et al. (2004), Zakikhany et al. (2007), Fan et al. (2005), de Jesus-Berrios et al. (2003), Missall et al. (2004), and Brown et al. (2009)
Superoxide dismutases (ROS detoxification)	Upregulated in *C. albicans, C. neoformans,* and *A. fumigatus.* Gene inactivation attenuates virulence in *C. albicans* and *C. neoformans* but not in *A. fumigatus.*	Hu et al. (2008), McDonagh et al. (2008), Morton et al. (2011), Fradin et al. (2003, 2005), Lorenz et al. (2004), Lambou et al. (2010), Missall et al. (2004), and Brown et al. (2009)

(continued)

Table 17.2 (continued)

Protein description	Gene expression modulation and functional analysis	References
Hydrophobin (cell surface protein)	Upregulated in *A. fumigatus, A. benhamiae,* and *T. rubrum.* Gene inactivation in *A. fumigatus* and *A. benhamiae* increases the susceptibility to the host immune response.	Cairns et al. (2010), Burmester et al. (2011), Heddergott et al. (2012), Aimanianda et al. (2009), and Lang et al. (2020)
HacA transcription factor (Unfolded protein response)	Gene inactivation in *A. fumigatus* and *T. rubrum* leads to attenuation in virulence traits and increases susceptibility to antifungal agents	Richie et al. (2009) and Bitencourt et al. (2020)
StuA transcription factor (APSES-family of the transcriptional regulators)	Upregulated in *T. rubrum*. Gene inactivation in *A. benhamiae* and *T. rubrum* impaired the growth on host molecules.	Krober et al. (2017) and Lang et al. (2020)
RlmA transcription factor (MPK1 mitogen-activated protein kinase pathway)	Upregulated in *A. fumigatus*. Gene inactivation in *A. fumigatus* decreases pathogenicity in mice.	Liu et al. (2021)

leakage of intracellular contents and fungal cell death. AMB also induces oxidative damage to cellular membranes via the generation of ROS. Some antifungal drugs target proteins involved in the ergosterol biosynthetic pathway (Martinez-Rossi et al. 2008) (Fig. 17.1). Azoles are the most commonly used class of antifungal drugs in clinical treatment and include ketoconazole, itraconazole, fluconazole, and voriconazole. They inhibit the activity of the enzyme cytochrome P450 lanosterol 14-α demethylase (ERG11), which is responsible for the oxidative removal of the 14α-methyl group of lanosterol, an essential step in ergosterol biosynthesis. Azoles are first-line agents for the treatment of candidiasis, but their frequent use can result in resistance due to their fungistatic mechanism of action. Terbinafine (TRB) is another antifungal drug that belongs to the allylamine class and is most effective against dermatophytes. It inhibits ergosterol biosynthesis by inhibiting the enzyme squalene epoxidase (ERG1), responsible for converting squalene to lanosterol. Inhibition of ERG1 decreases the production of ergosterol and increases the accumulation of squalene to toxic levels (Sagatova 2021) (Fig. 17.1).

Other antifungal drugs target DNA/RNA metabolism. Flucytosine is a cytosine analog that was first used as an antitumor agent. It also exhibits antifungal properties. Flucytosine is transported to the cytoplasm of fungal cells through cytosine permease; in the cytoplasm, cytosine deaminase converts it to 5-fluorouracil, which blocks protein and DNA synthesis. When phosphorylated, 5-fluorouracil is incorporated into RNA, leading to miscoding and inhibition of protein synthesis.

Table 17.3 Mechanism of action and mechanisms underlying antifungal resistance in fungi

Drug	Mechanism of action	Putative resistance mechanisms and drug response	References
Acriflavine	Topoisomerase inhibition/DNA intercalation. Nonspecific cellular interactions	Drug efflux, stress response, oxidative stress, decreases virulence	Fachin et al. (2006), Paiao et al. (2007), Persinoti et al. (2014), and Martinez-Rossi et al. (2016)
Amphotericin B	Binds irreversibly to ergosterol, resulting in disruption of membrane integrity	Drug efflux, stress response	Yu et al. (2007b), Martins et al. (2016), Mendes et al. (2016), and Bitencourt et al. (2019b)
Caspofungin	(1,3)-β-D-glucan synthase inhibition (encoded by FKS1/FKS2)	Mutations in FKS genes, posttranscriptional regulation of cell wall biosynthesis, stress response	Imtiaz et al. (2012), Perlin (2015), Bitencourt et al. (2019b), and Kalem et al. (2021)
Fluconazole	Cytochrome P450 14 α-lanosterol demethylase inhibition	Drug efflux, stress response, alteration of the drug target	Cervelatti et al. (2006), Fachin et al. (2006), Paiao et al. (2007), and Shapiro et al. (2011)
5-Flucytosine	DNA synthesis and nuclear division inhibition	Decreased drug uptake, alteration in enzyme activity, cell wall remodeling	Costa et al. (2015)
Griseofulvin	Mitosis inhibition	Drug efflux, stress response	Fachin et al. (1996, 2001, 2006), Cervelatti et al. (2006), Paiao et al. (2007), and Martins et al. (2016)
Imazalil	Cytochrome P450 14 α-lanosterol demethylase inhibition	Drug efflux	Cervelatti et al. (2006) and Fachin et al. (2006)
Itraconazole	Cytochrome P450 14 α-lanosterol demethylase inhibition	Drug efflux, alteration of the drug target	Cervelatti et al. (2006), Fachin et al. (2006), and Shapiro et al. (2011)
Ketoconazole	Cytochrome P450 14 α-lanosterol demethylase inhibition	Drug efflux	Cervelatti et al. (2006), Fachin et al. (2006), and Wang et al. (2021)
Terbinafine	Squalene epoxidase inhibition (encoded by Erg1)	Drug efflux, stress response, mutations in Erg1, drug metabolism	Graminha et al. (2004), Osborne et al. (2005, 2006), Rocha et al. (2006), Fachin et al. (2006), Paiao et al. (2007), Martins et al. (2016), Martinez-Rossi et al. (2016), Yamada et al. (2017), Santos et al. (2018), Petrucelli et al. (2019), and Kano (2021)

(continued)

Table 17.3 (continued)

Drug	Mechanism of action	Putative resistance mechanisms and drug response	References
Tioconazole	Cytochrome P450 14 α-lanosterol demethylase inhibition	Drug efflux	Fachin et al. (1996, 2001, 2006)
Undecanoic acid	Nonspecific cellular interactions	Stress response, drug metabolism, oxidative stress, decreases virulence	Paiao et al. (2007), Mendes et al. (2018), and Rossi et al. (2021)

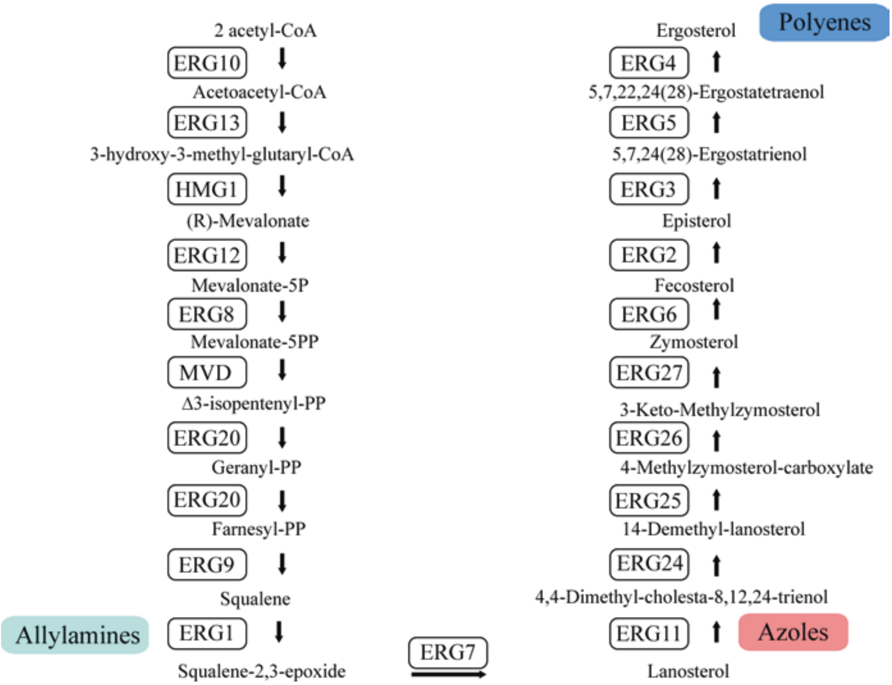


Fig. 17.1 Schematic representation of the ergosterol biosynthetic pathway

Furthermore, phosphorylated 5-fluorouracil can be converted into the deoxynucleoside form by uridine monophosphate pyrophosphorylase; thereafter, it inhibits the enzyme thymidylate synthetase and consequently disrupts DNA synthesis (Vermes et al. 2000; Billmyre et al. 2020). Griseofulvin, another antifungal drug, interacts with microtubules affecting the mitotic spindle formation, thereby inhibiting the mitosis in fungi. This drug serves as a fungistatic agent against dermatophytes. However, griseofulvin is not effective against dimorphic fungi, yeast, or chromomycosis-causing agents (Gupta et al. 2018).

The fungal cell wall is a specific target of antifungal drugs since it is absent from mammalian cells. Caspofungin was the first compound to target the fungal cell wall and was approved for clinical use in 2001. It is a member of the echinocandin class, which inhibits the enzyme (1,3)-β-D-glucan synthases (FKS1 and FKS2), thus preventing the synthesis of (1,3)-β-D-glucan and disrupting cell wall biosynthesis. In addition to caspofungin, two other echinocandins, micafungin and anidulafungin, are commercially available. These drugs are only available as intravenous infusions and are indicated to treat invasive aspergillosis and candidiasis. They have fungicidal activity against most *Candida* species and fungistatic activity against *Aspergillus* species. Although most fungal species encode orthologs of FKS1 and FKS2, echinocandins are not effective against *Zygomycetes* spp., *C. neoformans*, or *Fusarium* spp. (Perlin 2015; Kalem et al. 2021).

Transcriptome analyses have been used to evaluate the responses of pathogenic fungi, such as *C. albicans, A. fumigatus*, and *T. rubrum*, to several antifungal drugs, including azoles, polyenes, terbinafine, undecanoic acid, and echinocandins (Yu et al. 2007b; da Silva Ferreira et al. 2006; Gautam et al. 2008; Diao et al. 2009; Zhang et al. 2009; Peres et al. 2010b; Mendes et al. 2018; Cervelatti et al. 2006; Liu et al. 2005; Paiao et al. 2007). These studies revealed that the modulation of genes in the ergosterol biosynthetic pathway varies significantly among species and drugs. Although caspofungin and flucytosine do not primarily target the ergosterol biosynthetic pathway, they elicited the upregulation of some ergosterol biosynthetic genes in *C. albicans* (Liu et al. 2005). In response to ketoconazole, *C. albicans* upregulated genes involved in the biosynthesis of ergosterol, lipids, and fatty acids. Ketoconazole also induced the expression of the major transporter genes *CDR1* and *CDR2* (Liu et al. 2005). Similarly, in response to ketoconazole, *T. rubrum* upregulated genes involved in the metabolism of lipids, fatty acids, and sterols, as well as the multidrug-resistance gene encoding ABC1, which is a homolog of *C. albicans* CDR1 (Yu et al. 2007a). Transcriptome sequencing revealed that ketoconazole may also change cell membrane permeability, destroy the cell wall, and inhibit mitosis in *Microsporum canis* (Wang et al. 2021).

In response to AMB, *C. albicans* downregulated genes related to ergosterol biosynthesis and upregulated genes related to cell stress, including those encoding nitric oxide oxidoreductase (YHB1), catalase 1 (CTA1), aldehyde oxidase 1 (AOX1), and superoxide dismutase 2 (SOD2) (Liu et al. 2005). *A. fumigatus* exposed to AMB upregulated *erg11* and downregulated *erg6*. Besides, it modulated genes involved in cell stress, transport, oxidative phosphorylation, nucleotide metabolism, cell cycle control, and protein metabolism. Moreover, in response to the oxidative damage caused by AMB exposure, *A. fumigatus* overexpressed several genes encoding antioxidant enzymes, such as Mn-SOD, catalase, the thiol-specific antioxidant protein LsfA, glutathione S-transferase (GST), and thioredoxin. *A. fumigatus* downregulated ergosterol biosynthetic genes in response to AMB, possibly in an attempt to use alternate sterols or sterol intermediates in the cell membrane (Gautam et al. 2008). *C. albicans* exposed to caspofungin induced the expression of genes encoding cell wall maintenance proteins, including a target of caspofungin (the β-1,3-glucan synthase subunit homolog to FKS3), a pH-regulated glucan-remodeling enzyme (PHR1), extracellular matrix proteins (ECM21 and ECM33), and a putative fatty acid elongation enzyme (FEN12). Interestingly, *fen12* was upregulated in response to caspofungin and downregulated in response to AMB. In response to flucytosine, *C. albicans* upregulated the *CDC21* gene, which encodes thymidylate synthetase. This enzyme is the target of flucytosine and is associated with DNA synthesis; therefore, its upregulation may prevent fungal death. Other upregulated genes include those involved in purine and pyrimidine biosynthesis, such as YNK1, a nucleoside diphosphate kinase, and FUR1, an uracil phosphoribosyltransferase (Liu et al. 2005).

Terbinafine is commonly used to treat dermatophytosis. Exposure of *T. rubrum* to TRB decreased the expression of genes related to ergosterol biosyntheses, such as *erg2*, *erg4*, *erg24*, and *erg25*, and increased the expression of genes involved in lipid metabolism. Although TRB primary target is squalene epoxidase (ERG1), *T. rubrum* did not differentially express *erg1* after exposure to TRB. It did, however, upregulate multidrug-resistance (MDR) genes, including *mdr1* and *mdr2* (Zhang et al. 2009). Indeed, MDR2 is associated with drug susceptibility. Overexpression of *mdr*2 likely causes the efflux of TRB, since deletion of *mdr*2 increased dermatophyte susceptibility to TRB (Fachin et al. 2006). Interestingly, *in T. interdigitale*, the transcription of *mdr4* was downregulated in the Δ*mdr2* mutant challenged with amphotericin B or terbinafine, indicating that the transcription of *mdr4* is dependent on the function of *mdr2* in response to these drugs. However, when the Δ*mdr2* mutant was challenged with griseofulvin, the high expression of the *mdr4* gene seemed to compensate for the inactivation of the *mdr2* gene. These results suggest that these ABC transporter genes act synergistically, and they may compensate for one another when challenged with antifungal drugs (Martins et al. 2016; Martins et al. 2019). These results also indicate the existence of a network interaction responsible for the failure of antifungal therapeutics. An intriguing mechanism of resistance to TRB in *A. nidulans* (Graminha et al. 2004) and *T. rubrum* (Santos et al. 2018) involves the *salA* gene, which encodes a salicylate 1-monooxygenase. TRB contains a naphthalene nucleus in its molecular structure that might be degraded by salicylate 1-monooxygenase, an enzyme in the naphthalene degradation pathway in *Pseudomonas* (Bosch et al. 2000).

The emergence of resistant strains is an important obstacle to effective antifungal therapy. Azoles are the first-line treatment for many fungal infections; however, their use may select for azole-resistant mutants. Several mechanisms contribute to drug resistance, including alteration of the drug target, increased drug efflux, and increased cellular stress responses. Both mutations in and overexpression of the ergosterol biosynthesis gene *erg11*/*cyp51* confer resistance to azoles in *C. albicans* and *A. fumigatus*. For instance, one mutation causes the synthesis of an alternative protein insensitive to azoles and diminishes drug efficacy. At least 12 different point mutations in *erg11* have been identified in azole-resistant clinical isolates of *C. albicans* (Shapiro et al. 2011; Rosam et al. 2020). Overexpression of efflux pumps is associated with antifungal resistance in *C. albicans*. CDR1 and CDR2 confer resistance to multiple azoles, while MDR1 confers fluconazole resistance (White et al. 2002). Similarly, azole-resistant clinical isolates of *C. glabrata* have been shown to overexpress genes encoding CDR1 and CDR2 as well as SNQ2, another

ATP-binding cassette ABC transporter (Sanguinetti et al. 2005). In response to azoles and other structurally distinct drugs, dermatophytes overexpressed *mdr1* and *mdr2*, which encode ABC transporters (Cervelatti et al. 2006; Fachin et al. 2006). The zinc cluster transcription factor TAC1 regulates genes encoding the ABC transporters CDR1 and CDR2 in azole-resistant *C. albicans*, and deletion of *TAC1* gene prevented the upregulation of *cdr* genes (Coste et al. 2004). Furthermore, ChIP-chip experiments demonstrated that TAC1 directly binds to the promoter region of several genes, including CDR1, CDR2 (Liu et al. 2007).

Genome-wide expression analysis of resistant clinical isolates of *C. albicans* identified a transcription factor that was upregulated in coordination with MDR1. This gene encodes the multidrug resistance regulator MRR1, a zinc cluster transcription factor, and the main regulator of *MDR1* expression. Gain-of-function mutations in *MRR1* are responsible for overexpression of *MDR1* and are associated with fluconazole resistance in *C. albicans* (Morschhauser et al. 2007). In addition to regulating *MDR1* expression, MRR1 regulated at least 14 other genes that may also contribute to fluconazole resistance. These genes encoded mainly oxidoreductases. Notably, MRR1 does not target CDR1 or CDR2. Overall, large-scale transcriptional analyses have identified several genes associated with drug response and resistance in pathogenic fungi (Morschhauser et al. 2007). RNA-seq analyses were performed on two isogenic *C. albicans* strains that differed only in fluconazole resistance. These studies identified novel genes associated with azole resistance, including the transcription factor CZF1, which is involved in the hyphal transition and the white/opaque switch. Inactivation of CZF1 increased the susceptibility to fluconazole and unrelated antifungal drugs, such as TRB and anisomycin. Furthermore, the CZF1 mutant strain displayed increased resistance to the cell wall-disrupting agent Congo red. The mutant also overexpressed the gene encoding β 1,3-glucan synthase (GLS1), suggesting that CZF1 represses β-glucan synthesis and regulates cell wall integrity (Dhamgaye et al. 2012).

The transcription profile of a *C. auris* isolate susceptible to AMB and voriconazole showed upregulation of the expression of 39 genes in response to AMB, 21 in response to voriconazole, and 14 being upregulated in response to both drugs (Munoz et al. 2018). AMB-responsive genes included those involved in arginine synthesis (ARG1/ARG3), ergosterol biosynthesis (ERG24), fatty acid metabolism (FAS1/FAS2), GPI-linked surface proteins (PGA7 and RBT5), and iron transporters, such as SIT1. In response to both drugs, there was an upregulation of the expression of transmembrane transport and iron transport-related genes, such as the high-affinity iron transporter FTH1, ferric reductase, glucose transporter, N-acetylglucosamine transporter, and OP1-like oligopeptide transporter. In a *C. auris*-resistant strain to AMB and voriconazole, AMB-responsive genes were enriched in translation and transcription processes and the sterol biosynthetic pathway. Voriconazole response induced the expression of genes related to RNA processing and transcription. Furthermore, both strains induced the expression of the genes *ARG1*, *CSA1*, *MET15*, and OPT1-like transporter in response to AMB. The authors also showed an intrinsic expression profile of polyene resistance genes, such as D-xylulose reductase, phosphoenolpyruvate carboxykinase, and several transporters and stress response genes in the resistant strain (Munoz et al. 2018).

In pathogenic fungi, mitochondrial dysfunction has been associated with altered susceptibility to antifungal drugs. In *C. albicans*, inhibition or mutation of the mitochondrial complex I (CI) increased susceptibly to fluconazole even in resistant clinical isolates. Transcriptional analysis was performed on the Δ*goa1* and Δ*ndh51* mutant strains, which are associated with CI-induced susceptibility to fluconazole. GOA1 is required for the function of the electron transport chain, and the Δ*goa1* mutant accumulates ROS, undergoes apoptosis, and is avirulent. *Ndh51* encodes a 51-kDa subunit of the NADH dehydrogenase of the electron transport chain, and the Δ*ndh51* mutant exhibits defects in morphogenesis. RNA-seq analyses of these strains demonstrated downregulation of transporters, including the CDR1/CDR2 efflux pumps but not MDR1. Genes related to ergosterol biosynthesis were downregulated in the Δ*ndh51* mutant. In contrast, genes associated with peroxisomes, gluconeogenesis, β-oxidation, and mitochondria were downregulated in the Δ*goa1* mutant (Sun et al. 2013). NDH51 is conserved among eukaryotes, including mammals; nevertheless, GOA1 is conserved only in some *Candida* species. Therefore, fungi-specific mitochondrial genes may be targets for the development of novel antifungal drugs. Indeed, acriflavine, an acridine derivative with antibacterial, antifungal, antiviral, and antiparasitic properties, induces the overexpression of genes involved in the mitochondrial electron transport chain of *T. rubrum* (Segato et al. 2008). Transcriptomic analysis of the effect of acriflavine on *T. rubrum* showed that the expression of genes involved in cellular detoxification was upregulated, protecting the cell against oxidative stress and reactive oxygen species. Furthermore, this drug interferes with the establishment and maintenance of the fungal infection (Persinoti et al. 2014).

Interestingly, chemical inhibition of fungal HSP90 improved the activity of azoles and echinocandins against *C. albicans* and echinocandins against *A. fumigatus* (Cowen 2009). Inhibition of HPS90 prevents the stress-response cascade mediated by calcineurin, which is normally activated in response to antifungal drugs. Blunting of the stress-response cascade enhances the fungicidal effects, leading to cell death. The development of an inhibitor selective for fungal HSP90 and inactive against human HSP90 has been challenging. Nevertheless, HSP90 is a promising target for the treatment of resistant fungal diseases and may combat the emergence of drug resistance (Cowen 2009; Martinez-Rossi et al. 2016). Additionally, chemical inhibition of Hsp90 of *T. rubrum* increased the susceptibility to itraconazole and micafungin and decreased its ability to grow on human nail fragments. These results suggest the role of Hsp90 in the pathogenicity and drug susceptibility in *T. rubrum* (Jacob et al. 2015), reinforcing its potential as a target for the treatment of fungal infections.

In addition to the emergence of drug-resistant strains, another major clinical problem is the formation of microbial biofilms. Biofilms possess specific traits as compared to planktonic cells, such as intrinsic resistance to drugs. In immunocompromised individuals, both *C. albicans* and *A. fumigatus* can form biofilms on implanted medical devices, such as catheters, and cause persistent infections. In particular, biofilms have decreased susceptibility to antifungal drugs. To understand their mechanisms of resistance, mature biofilm cells were exposed to fluconazole, AMB, and caspofungin. Fluconazole exposure did not significantly alter gene expression, and AMB exposure resulted in only minor alterations in gene expression. On the other hand, biofilms exposed to caspofungin underwent more pronounced alteration in gene expression, including the upregulation of several genes associated with biofilm formation, such as ALS3, a cell wall adhesin, the transcription factor TEC1, and genes associated with cell wall remodeling (Vediyappan et al. 2010). Furthermore, AMB and fluconazole bind to the extracellular matrix of the biofilm, which is comprised of β-glucans; such binding inhibits effective drug action (Vediyappan et al. 2010). An RNA-seq analysis compared the transcriptional profile of an *A. fumigatus* biofilm to that of planktonic cells. Thousands of genes were differentially expressed between the biofilm and planktonic cells. Specifically, the biofilm exhibited an upregulation of secondary metabolism genes, cell wall-related genes, sterol biosynthetic genes (e.g., *erg11*), transporters associated with antifungal resistance (MDR1, MDR2, and MDR4), and hydrophobins, which are associated with the structural organization of biofilms (Gibbons et al. 2012). The complex gene network involved in biofilm formation is consistent with the fact that *C. albicans* can form biofilms in different niches, such as the bloodstream, oral cavity, or medical devices serving as reservoirs of drug-resistant cells (Mamouei et al. 2021; Li et al. 2021). This highlights the challenge inherent to treating these infections as well as the importance of searching for new antifungal targets.

Furthermore, posttranscriptional regulation has been described as a fine adjustment for some fungi to adapt to antifungal exposure. This phenomenon occurs in *N. crassa* through alternative splicing of pre-mRNA transcripts of genes encoding asparagine synthetase 2, C6-zinc-finger regulator, and farnesyltransferase in response to amphotericin B and ketoconazole (Mendes et al. 2016). Additionally, RNA-Seq analysis of *T. rubrum* exposed to undecanoic acid revealed alternative splicing in several genes, including *hsp*s (Mendes et al. 2018; Neves-da-Rocha et al. 2019). These results show the complexity of the metabolic modulation triggered by antifungal signaling.

In conclusion, analyses of the transcriptional changes in response to cytotoxic drugs have identified genes with known biological functions, suggesting novel effects of antifungal drugs. Besides, some of the drug-responsive genes are shared across multiple classes of antifungal agents in *C. albicans* (Liu et al. 2005), dermatophytes (Peres et al. 2010b; Persinoti et al. 2014; Mendes et al. 2018; Fachin et al. 2006; Paiao et al. 2007), and other fungi. Nonspecific responses to stress are also known that allow fungi to adapt to several drugs and environmental challenges, highlighting the broad range of fungal responses to cope with stress.

17.5 Concluding Remarks

The pathogenesis of fungal infections involves gene expression changes and metabolic pathways, which enable fungal invasion, survival, and dissemination. At the same time, fungi elicit host responses aimed at eliminating the pathogen.

Genome-wide transcriptional profiling has identified the molecular responses of both host and pathogen during the interaction. It has provided insights into the adaptive responses that occur during the establishment of infection, antifungal resistance, and exposure. The combination of large-scale transcriptomic analysis and systems biology approaches has enabled the development of regulatory molecular models that can aid in the assessment of dynamic behaviors of host–pathogen interactions and elucidate the pathogenesis of human mycoses. These regulatory models have been validated through reverse-genetic approaches by evaluating the physiological behavior of the knockout strains under *in vitro*, *ex vivo*, and *in vivo conditions*. Furthermore, transcriptomics is a valuable source of data on gene expression regulation, gene structure and function, and information regarding the mechanisms of fungal responses and resistance to drugs. These insights will further support the development of novel therapeutic approaches to prevent and control fungal infections.

Acknowledgments This work was supported by grants from Brazilian funding agencies FAPESP (Grant No. 2019/22596-9), National Council for Scientific and Technological Development (CNPq) (Grants No. 305797/2017-4 and 304989/2017-7); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001; and Fundação de Apoio ao Ensino, Pesquisa e Assistência (FAEPA).

FAPESP supported the postdoctoral fellowships to N.P. (2009/08411-4), T.B. (2015/23435-8), G.P. (2012/22232-8 and 2013/19195-6), and E.L. (2011/08424-9).

References

Aimanianda V, Bayry J, Bozza S, Kniemeyer O, Perruccio K, Elluru SR, Clavaud C, Paris S, Brakhage AA, Kaveri SV, Romani L, Latge JP (2009) Surface hydrophobin prevents immune recognition of airborne fungal spores. Nature 460(7259):1117–1121

Alanio A, Delliere S, Fodil S, Bretagne S, Megarbane B (2020) Prevalence of putative invasive pulmonary aspergillosis in critically ill patients with COVID-19. Lancet Respir Med 8(6):e48–e49

Amich J, Vicentefranqueira R, Leal F, Calera JA (2010) *Aspergillus fumigatus* survival in alkaline and extreme zinc-limiting environments relies on the induction of a zinc homeostasis system encoded by the *zrfC* and *aspf2* genes. Eukaryot Cell 9 (3):424–437

Ballou ER, Avelar GM, Childers DS, Mackie J, Bain JM, Wagener J, Kastora SL, Panea MD, Hardison SE, Walker LA, Erwig LP, Munro CA, Gow NA, Brown GD, MacCallum DM, Brown AJ (2016) Lactate signalling regulates fungal beta-glucan masking and immune evasion. Nat Microbiol 2:16238

Barelle CJ, Priest CL, Maccallum DM, Gow NA, Odds FC, Brown AJ (2006) Niche-specific regulation of central metabolic pathways in a fungal pathogen. Cell Microbiol 8(6):961–971

Barker KS, Park H, Phan QT, Xu L, Homayouni R, Rogers PD, Filler SG (2008) Transcriptome profile of the vascular endothelial cell response to *Candida albicans*. J Infect Dis 198(2):193–202

Bedoya SK, Lam B, Lau K, Larkin J 3rd (2013) Th17 cells in immunity and autoimmunity. Clin Dev Immunol 2013:986789

Bielska E, May RC (2019) Extracellular vesicles of human pathogenic fungi. Curr Opin Microbiol 52:90–99

Billmyre RB, Applen Clancey S, Li LX, Doering TL, Heitman J (2020) 5-fluorocytosine resistance is associated with hypermutation and alterations in capsule biosynthesis in *Cryptococcus*. Nat Commun 11(1):127

Biondo C, Midiri A, Gambuzza M, Gerace E, Falduto M, Galbo R, Bellantoni A, Beninati C, Teti G, Leanderson T, Mancuso G (2008) IFN-alpha/beta signaling is required for polarization of cytokine responses toward a protective type 1 pattern during experimental cryptococcosis. J Immunol 181(1):566–573

Biondo C, Signorino G, Costa A, Midiri A, Gerace E, Galbo R, Bellantoni A, Malara A, Beninati C, Teti G, Mancuso G (2011) Recognition of yeast nucleic acids triggers a host-protective type I interferon response. Eur J Immunol 41(7):1969–1979

Bitencourt TA, Macedo C, Franco ME, Assis AF, Komoto TT, Stehling EG, Beleboni RO, Malavazi I, Marins M, Fachin AL (2016) Transcription profile of *Trichophyton rubrum* conidia grown on keratin reveals the induction of an adhesin-like protein gene with a tandem repeat pattern. BMC Genomics 17:249

Bitencourt TA, Rezende CP, Quaresemin NR, Moreno P, Hatanaka O, Rossi A, Martinez-Rossi NM, Almeida F (2018) Extracellular vesicles from the dermatophyte *Trichophyton interdigitale* modulate macrophage and keratinocyte functions. Front Immunol 9:2343

Bitencourt TA, Macedo C, Franco ME, Rocha MC, Moreli IS, Cantelli BAM, Sanches PR, Beleboni RO, Malavazi I, Passos GA, Marins M, Fachin AL (2019a) Trans-chalcone activity against *Trichophyton rubrum* relies on an interplay between signaling pathways related to cell wall integrity and fatty acid metabolism. BMC Genomics 20(1):411

Bitencourt TA, Oliveira FB, Sanches PR, Rossi A, Martinez-Rossi NM (2019b) The *prp4 kinase* gene and related spliceosome factor genes in *Trichophyton rubrum* respond to nutrients and antifungals. J Med Microbiol 68(4):591–599

Bitencourt TA, Lang EAS, Sanches PR, Peres NTA, Oliveira VM, Fachin AL, Rossi A, Martinez-Rossi NM (2020) HacA governs virulence traits and adaptive stress responses in *Trichophyton rubrum*. Front Microbiol 11:193

Bosch R, Garcia-Valdes E, Moore ER (2000) Complete nucleotide sequence and evolutionary significance of a chromosomally encoded naphthalene-degradation lower pathway from *Pseudomonas stutzeri* AN10. Gene 245(1):65–74

Brock M (2009) Fungal metabolism in host niches. Curr Opin Microbiol 12(4):371–376

Brown AJ, Haynes K, Quinn J (2009) Nitrosative and oxidative stress responses in fungal pathogenicity. Curr Opin Microbiol 12(4):384–391

Brown GD, Denning DW, Gow NA, Levitz SM, Netea MG, White TC (2012) Hidden killers: human fungal infections. Sci Transl Med 4(165):165rv113

Brown AJ, Budge S, Kaloriti D, Tillmann A, Jacobsen MD, Yin Z, Ene IV, Bohovych I, Sandai D, Kastora S, Potrykus J, Ballou ER, Childers DS, Shahana S, Leach MD (2014) Stress adaptation in a pathogenic fungus. J Exp Biol 217(Pt 1):144–155

Bruno M, Kersten S, Bain JM, Jaeger M, Rosati D, Kruppa MD, Lowman DW, Rice PJ, Graves B, Ma Z, Jiao YN, Chowdhary A, Renieris G, van de Veerdonk FL, Kullberg BJ, Giamarellos-Bourboulis EJ, Hoischen A, Gow NAR, Brown AJP, Meis JF, Williams DL, Netea MG (2020) Transcriptional and functional insights into the host immune response against the emerging fungal pathogen *Candida auris*. Nat Microbiol 5(12):1516–1531

Burmester A, Shelest E, Glockner G, Heddergott C, Schindler S, Staib P, Heidel A, Felder M, Petzold A, Szafranski K, Feuermann M, Pedruzzi I, Priebe S, Groth M, Winkler R, Li W, Kniemeyer O, Schroeckh V, Hertweck C, Hube B, White TC, Platzer M, Guthke R, Heitman J, Wostemeyer J, Zipfel PF, Monod M, Brakhage AA (2011) Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. Genome Biol 12(1):R7

Burstein VL, Beccacece I, Guasconi L, Mena CJ, Cervi L, Chiapello LS (2020) Skin immunity to dermatophytes: from experimental infection models to human disease. Front Immunol 11:605644

Cairns T, Minuzzi F, Bignell E (2010) The host-infecting fungal transcriptome. FEMS Microbiol Lett 307(1):1–11

Cambier L, Weatherspoon A, Defaweux V, Bagut ET, Heinen MP, Antoine N, Mignon B (2014) Assessment of the cutaneous immune response during *Arthroderma benhamiae* and *A. vanbreuseghemii* infection using an experimental mouse model. Br J Dermatol 170(3):625–633

Cavalheiro M, Teixeira MC (2018) Candida biofilms: threats, challenges, and promising strategies. Front Med (Lausanne) 5:28

Cervelatti EP, Fachin AL, Ferreira-Nozawa MS, Martinez-Rossi NM (2006) Molecular cloning and characterization of a novel ABC transporter gene in the human pathogen Trichophyton rubrum. Med Mycol 44(2):141–147

Chen Y, Toffaletti DL, Tenor JL, Litvintseva AP, Fang C, Mitchell TG, McDonald TR, Nielsen K, Boulware DR, Bicanic T, Perfect JR (2014) The *Cryptococcus neoformans* transcriptome at the site of human meningitis. MBio 5(1):e01087–e01013

Citiulo F, Jacobsen ID, Miramón P, Schild L, Brunke S, Zipfel P, Brock M, Hube B, Wilson D (2012) Candida albicans Scavenges Host Zinc via Pra1 during Endothelial Invasion. Plos Pathogens 8(6):e1002777. https://doi.org/10.1371/journal.ppat.1002777

Cornet M, Gaillardin C (2014) pH signaling in human fungal pathogens: a new target for antifungal strategies. Eukaryot Cell 13(3):342–352

Cortez KJ, Lyman CA, Kottilil S, Kim HS, Roilides E, Yang J, Fullmer B, Lempicki R, Walsh TJ (2006) Functional genomics of innate host defense molecules in normal human monocytes in response to *Aspergillus fumigatus*. Infect Immun 74(4):2353–2365

Costa C, Ponte A, Pais P, Santos R, Cavalheiro M, Yaguchi T, Chibana H, Teixeira MC (2015) New mechanisms of flucytosine resistance in *C. glabrata* unveiled by a chemogenomics analysis in S. cerevisiae. PLoS One 10(8):e0135110

Coste AT, Karababa M, Ischer F, Bille J, Sanglard D (2004) TAC1, transcriptional activator of CDR genes, is a new transcription factor involved in the regulation of *Candida albicans* ABC transporters CDR1 and CDR2. Eukaryot Cell 3(6):1639–1652

Cowen LE (2009) Hsp90 orchestrates stress response signaling governing fungal drug resistance. PLoS Pathog 5(8):e1000471

da Silva Ferreira ME, Malavazi I, Savoldi M, Brakhage AA, Goldman MH, Kim HS, Nierman WC, Goldman GH (2006) Transcriptome analysis of *Aspergillus fumigatus* exposed to voriconazole. Curr Genet 50(1):32–44

da Silva LG, Martins MP, Sanches PR, Peres NTA, Martinez-Rossi NM, Rossi A (2020) Saline stress affects the pH-dependent regulation of the transcription factor PacC in the dermatophyte *Trichophyton interdigitale*. Braz J Microbiol 51(4):1585–1591

de Jesus-Berrios M, Liu L, Nussbaum JC, Cox GM, Stamler JS, Heitman J (2003) Enzymes that counteract nitrosative stress promote fungal virulence. Curr Biol 13(22):1963–1968

d'Enfert C, Kaune AK, Alaban LR, Chakraborty S, Cole N, Delavy M, Kosmala D, Marsaux B, Frois-Martins R, Morelli M, Rosati D, Valentine M, Xie Z, Emritloll Y, Warn PA, Bequet F, Bougnoux ME, Bornes S, Gresnigt MS, Hube B, Jacobsen ID, Legrand M, Leibundgut-Landmann S, Manichanh C, Munro CA, Netea MG, Queiroz K, Roget K, Thomas V, Thoral C, Van den Abbeele P, Walker AW, Brown AJP (2020) The impact of the fungus-host-microbiota interplay upon *Candida albicans* infections: current knowledge and new perspectives. FEMS Microbiol Rev 45

Deng W, Liang P, Zheng Y, Su Z, Gong Z, Chen J, Feng P, Chen J (2020) Differential gene expression in HaCaT cells may account for the various clinical presentation caused by anthropophilic and geophilic dermatophytes infections. Mycoses 63(1):21–29

de-Souza-Silva CM, Hurtado FA, Tavares AH, de Oliveira GP Jr, Raiol T, Nishibe C, Agustinho DP, Almeida NF, Walter M, Nicola AM, Bocca AL, Albuquerque P, Silva-Pereira I (2020) Transcriptional remodeling patterns in murine dendritic cells infected with *Paracoccidioides brasiliensis*: more is not necessarily better. J Fungi (Basel) 6(4)

Dhamgaye S, Bernard M, Lelandais G, Sismeiro O, Lemoine S, Coppee JY, Le Crom S, Prasad R, Devaux F (2012) RNA sequencing revealed novel actors of the acquisition of drug resistance in *Candida albicans*. BMC Genomics 13:396

Diao YJ, Zhao R, Deng XM, Leng WC, Peng JP, Jin Q (2009) Transcriptional profiles of *Trichophyton rubrum* in response to itraconazole. Med Mycol 47(3):237–247

Du H, Bing J, Hu T, Ennis CL, Nobile CJ, Huang G (2020) *Candida auris*: epidemiology, biology, antifungal resistance, and virulence. PLoS Pathog 16(10):e1008921

Fachin AL, Maffei CM, Martinez-Rossi NM (1996) In vitro susceptibility of *Trichophyton rubrum* isolates to griseofulvin and tioconazole. Induction and isolation of a resistant mutant to both antimycotic drugs. Mutant of *Trichophyton rubrum* resistant to griseofulvin and tioconazole. Mycopathologia 135(3):141–143

Fachin AL, Contel EP, Martinez-Rossi NM (2001) Effect of sub-MICs of antimycotics on expression of intracellular esterase of *Trichophyton rubrum*. Med Mycol 39(1):129–133

Fachin AL, Ferreira-Nozawa MS, Maccheroni W, Martinez-Rossi NM (2006) Role of the ABC transporter TruMDR2 in terbinafine, 4-nitroquinoline N-oxide and ethidium bromide susceptibility in *Trichophyton rubrum*. J Med Microbiol 55(Pt 8):1093–1099

Fan W, Kraus PR, Boily MJ, Heitman J (2005) *Cryptococcus neoformans* gene expression during murine macrophage infection. Eukaryot Cell 4 (8):1420–1433

Ferreira-Nozawa MS, Silveira HCS, Ono CJ, Fachin AL, Rossi A, Martinez-Rossi NM (2006) The pH signaling transcription factor PacC mediates the growth of *Trichophyton rubrum* on human nail in vitro. Med Mycol 44(7):641–645

Firat YH, Simanski M, Rademacher F, Schroder L, Brasch J, Harder J (2014) Infection of keratinocytes with *Trichophytum rubrum* induces epidermal growth factor-dependent RNase 7 and human beta-defensin-3 expression. PLoS One 9(4):e93941

Fradin C, Kretschmar M, Nichterlein T, Gaillardin C, d'Enfert C, Hube B (2003) Stage-specific gene expression of *Candida albicans* in human blood. Mol Microbiol 47(6):1523–1543

Fradin C, De Groot P, MacCallum D, Schaller M, Klis F, Odds FC, Hube B (2005) Granulocytes govern the transcriptional response, morphology and proliferation of *Candida albicans* in human blood. Mol Microbiol 56(2):397–415

Fradin C, Mavor AL, Weindl G, Schaller M, Hanke K, Kaufmann SH, Mollenkopf H, Hube B (2007) The early transcriptional response of human granulocytes to infection with *Candida albicans* is not essential for killing but reflects cellular communications. Infect Immun 75(3):1493–1501

Frazzitta AE, Vora H, Price MS, Tenor JL, Betancourt-Quiroz M, Toffaletti DL, Cheng N, Perfect JR (2013) Nitrogen source-dependent capsule induction in human-pathogenic cryptococcus species. Eukaryot Cell 12(11):1439–1450

Gautam P, Shankar J, Madan T, Sirdeshmukh R, Sundaram CS, Gade WN, Basir SF, Sarma PU (2008) Proteomic and transcriptomic analysis of *Aspergillus fumigatus* on exposure to amphotericin B. Antimicrob Agents Chemother 52(12):4220–4227

Gibbons JG, Beauvais A, Beau R, McGary KL, Latge JP, Rokas A (2012) Global transcriptome changes underlying colony growth in the opportunistic human pathogen *Aspergillus fumigatus*. Eukaryot Cell 11(1):68–78

Gomez P, Hackett TL, Moore MM, Knight DA, Tebbutt SJ (2011) Functional genomics of human bronchial epithelial cells directly interacting with conidia of *Aspergillus fumigatus*. BMC Genomics 11:358

Gonzalez Segura G, Cantelli BA, Peronni K, Rodrigo Sanches P, Komoto TT, Rizzi E, Beleboni RO, Junior W, Martinez-Rossi NM, Marins M, Fachin AL (2020) Cellular and molecular response of macrophages THP-1 during co-culture with inactive *Trichophyton rubrum* Conidia. J Fungi (Basel) 6(4)

Graminha MA, Rocha EM, Prade RA, Martinez-Rossi NM (2004) Terbinafine resistance mediated by salicylate 1-monooxygenase in *Aspergillus nidulans*. Antimicrob Agents Chemother 48(9):3530–3535

Gupta AK, Mays RR, Versteeg SG, Piraccini BM, Shear NH, Piguet V, Tosti A, Friedlander SF (2018) *Tinea capitis* in children: a systematic review of management. J Eur Acad Dermatol Venereol 32(12):2264–2274

Heddergott C, Bruns S, Nietzsche S, Leonhardt I, Kurzai O, Kniemeyer O, Brakhage AA (2012) The *Arthroderma benhamiae* hydrophobin HypA mediates hydrophobicity and influences recognition by human immune effector cells. Eukaryot Cell 11(5):673–682

Hu G, Cheng PY, Sham A, Perfect JR, Kronstad JW (2008) Metabolic adaptation in *Cryptococcus neoformans* during early murine pulmonary infection. Mol Microbiol 69(6):1456–1475

Huang MY, Woolford CA, May G, McManus CJ, Mitchell AP (2019) Circuit diversification in a biofilm regulatory network. PLoS Pathog 15(5):e1007787

Ibrahim-Granet O, Dubourdeau M, Latge JP, Ave P, Huerre M, Brakhage AA, Brock M (2008) Methylcitrate synthase from *Aspergillus fumigatus* is essential for manifestation of invasive aspergillosis. Cell Microbiol 10(1):134–148

Idnurm A, Giles SS, Perfect JR, Heitman J (2007) Peroxisome function regulates growth on glucose in the basidiomycete fungus *Cryptococcus neoformans*. Eukaryot Cell 6(1):60–72

Imtiaz T, Lee KK, Munro CA, MacCallum DM, Shankland GS, Johnson EM, MacGregor MS, Bal AM (2012) Echinocandin resistance due to simultaneous FKS mutation and increased cell wall chitin in a *Candida albicans* bloodstream isolate following brief exposure to caspofungin. J Med Microbiol 61(Pt 9):1330–1334

Inglis DO, Berkes CA, Hocking Murray DR, Sil A (2010) Conidia but not yeast cells of the fungal pathogen *Histoplasma capsulatum* trigger a type I interferon innate immune response in murine macrophages. Infect Immun 78(9):3871–3882

Jacob TR, Peres NT, Martins MP, Lang EA, Sanches PR, Rossi A, Martinez-Rossi NM (2015) Heat shock protein 90 (Hsp90) as a molecular target for the development of novel drugs against the dermatophyte *Trichophyton rubrum*. Front Microbiol 6:1241

Johns LE, Goldman GH, Ries LNA, Brown NA (2021) Nutrient sensing and acquisition in fungi: mechanisms promoting pathogenesis in plant and human hosts. Fungal Biol Rev 36:1–14

Kalem MC, Subbiah H, Leipheimer J, Glazier VE, Panepinto JC (2021) Puf4 mediates post-transcriptional regulation of cell wall biosynthesis and caspofungin resistance in *Cryptococcus neoformans*. MBio 12(1)

Kano R (2021) ATP-binding Cassette (ABC) transporter proteins in highly terbinafine-resistant strains of T*richophyton indotineae* (Former species name: *Trichophyton interdigitale)*. Med Mycol J 62(1):21–25

Kean R, Delaney C, Sherry L, Borman A, Johnson EM, Richardson MD, Rautemaa-Richardson R, Williams C, Ramage G (2018) Transcriptome assembly and profiling of *Candida auris* reveals novel insights into biofilm-mediated resistance. mSphere 3(4)

Kim HS, Choi EH, Khan J, Roilides E, Francesconi A, Kasai M, Sein T, Schaufele RL, Sakurai K, Son CG, Greer BT, Chanock S, Lyman CA, Walsh TJ (2005) Expression of genes encoding innate host defense molecules in normal human monocytes in response to *Candida albicans*. Infect Immun 73(6):3714–3724

Kretschmer M, Wang J, Kronstad JW (2012) Peroxisomal and mitochondrial beta-oxidation pathways influence the virulence of the pathogenic fungus *Cryptococcus neoformans*. Eukaryot Cell 11(8):1042–1054

Krober A, Etzrodt S, Bach M, Monod M, Kniemeyer O, Staib P, Brakhage AA (2017) The transcriptional regulators SteA and StuA contribute to keratin degradation and sexual reproduction of the dermatophyte *Arthroderma benhamiae*. Curr Genet 63(1):103–116

Lambou K, Lamarre C, Beau R, Dufour N, Latge JP (2010) Functional analysis of the superoxide dismutase family in *Aspergillus fumigatus*. Mol Microbiol 75(4):910–923

Lang EAS, Bitencourt TA, Peres NTA, Lopes L, Silva LG, Cazzaniga RA, Rossi A, Martinez-Rossi NM (2020) The stuA gene controls development, adaptation, stress tolerance, and virulence of the dermatophyte *Trichophyton rubrum*. Microbiol Res 241:126592

Li H, Li Y, Sun T, Du W, Li C, Suo C, Meng Y, Liang Q, Lan T, Zhong M, Yang S, Niu C, Li D, Ding C (2019) Unveil the transcriptional landscape at the *Cryptococcus*-host axis in mice and nonhuman primates. PLoS Negl Trop Dis 13(7):e0007566

Li P, Seneviratne CJ, Luan Q, Jin L (2021) Proteomic analysis of caspofungin-induced responses in planktonic cells and biofilms of *Candida albicans*. Front Microbiol 12:639123

Lim CS, Rosli R, Seow HF, Chong PP (2011) Transcriptome profiling of endothelial cells during infections with high and low densities of *C. albicans* cells. Int J Med Microbiol 301(6):536–546

Liu TT, Lee RE, Barker KS, Lee RE, Wei L, Homayouni R, Rogers PD (2005) Genome-wide expression profiling of the response to azole, polyene, echinocandin, and pyrimidine antifungal agents in *Candida albicans*. Antimicrob Agents Chemother 49(6):2226–2236

Liu TT, Znaidi S, Barker KS, Xu L, Homayouni R, Saidane S, Morschhauser J, Nantel A, Raymond M, Rogers PD (2007) Genome-wide expression and location analyses of the *Candida albicans* Tac1p regulon. Eukaryot Cell 6(11):2122–2138

Liu H, Xu W, Bruno VM, Phan QT, Solis NV, Woolford CA, Ehrlich RL, Shetty AC, McCraken C, Lin J, Bromley MJ, Mitchell AP, Filler SG (2021) Determining *Aspergillus fumigatus* transcription factor expression and function during invasion of the mammalian lung. PLoS Pathog 17(3):e1009235

Livonesi MC, Souto JT, Campanelli AP, Maffei CM, Martinez R, Rossi MA, Da Silva JS (2008) Deficiency of IL-12p40 subunit determines severe paracoccidioidomycosis in mice. Med Mycol 46(7):637–646

Lopes L, Bitencourt TA, Lang EAS, Sanches PR, Peres NTA, Rossi A, Martinez-Rossi NM (2019) Genes coding for LysM domains in the dermatophyte *Trichophyton rubrum*: a transcription analysis. Med Mycol 58(3):372–379

Lorenz MC, Fink GR (2001) The glyoxylate cycle is required for fungal virulence. Nature 412(6842):83–86

Lorenz MC, Bender JA, Fink GR (2004) Transcriptional response of *Candida albicans* upon internalization by macrophages. Eukaryot Cell 3(5):1076–1087

Luberto C, Martinez-Marino B, Taraskiewicz D, Bolanos B, Chitano P, Toffaletti DL, Cox GM, Perfect JR, Hannun YA, Balish E, Del Poeta M (2003) Identification of App1 as a regulator of phagocytosis and virulence of *Cryptococcus neoformans*. J Clin Invest 112(7):1080–1094

Lupo P, Chang YC, Kelsall BL, Farber JM, Pietrella D, Vecchiarelli A, Leon F, Kwon-Chung KJ (2008) The presence of capsule in *Cryptococcus neoformans* influences the gene expression profile in dendritic cells during interaction with the fungus. Infect Immun 76(4):1581–1589

Mamouei Z, Singh S, Lemire B, Gu Y, Alqarihi A, Nabeela S, Li D, Ibrahim A, Uppuluri P (2021) An evolutionarily diverged mitochondrial protein controls biofilm growth and virulence in *Candida albicans*. PLoS Biol 19(3):e3000957

Maranhão FCA, Paião FG, Martinez-Rossi NM (2007) Isolation of transcripts over-expressed in human pathogen *Trichophyton rubrum* during growth in keratin. Microb Pathog 43(4):166–172

Martinez-Rossi NM, Peres NT, Rossi A (2008) Antifungal resistance mechanisms in dermatophytes. Mycopathologia 166(5-6):369–383

Martinez-Rossi NM, Persinoti GF, Peres NTA, Rossi A (2012) Role of pH in the pathogenesis of dermatophytoses. Mycoses 55(5):381–387

Martinez-Rossi NM, Jacob TR, Sanches PR, Peres NT, Lang EA, Martins MP, Rossi A (2016) Heat shock proteins in dermatophytes: current advances and perspectives. Curr Genomics 17(2):99–111

Martinez-Rossi NM, Peres NT, Rossi A (2017) Pathogenesis of dermatophytosis: sensing the host tissue. Mycopathologia 182(1-2):215–227

Martinez-Rossi NM, Bitencourt TA, Peres NTA, Lang EAS, Gomes EV, Quaresemin NR, Martins MP, Lopes L, Rossi A (2018) Dermatophyte resistance to antifungal drugs: mechanisms and prospectus. Front Microbiol 9:1108

Martins MP, Franceschini ACC, Jacob TR, Rossi A, Martinez-Rossi NM (2016) Compensatory expression of multidrug-resistance genes encoding ABC transporters in dermatophytes. J Med Microbiol 65(7):605–610

Martins MP, Rossi A, Sanches PR, Martinez-Rossi NM (2019) Differential expression of multidrug-resistance genes in *Trichophyton rubrum*. J I OMICS 9(2):65–69

Martins MP, Martinez-Rossi NM, Sanches PR, Rossi A (2020a) The PAC-3 transcription factor critically regulates phenotype-associated genes in *Neurospora crassa*. Genet Mol Biol 43(3):e20190374

Martins MP, Rossi A, Sanches PR, Bortolossi JC, Martinez-Rossi NM (2020b) Comprehensive analysis of the dermatophyte *Trichophyton rubrum* transcriptional profile reveals dynamic metabolic modulation. Biochem J 477(5):873–885

McDonagh A, Fedorova ND, Crabtree J, Yu Y, Kim S, Chen D, Loss O, Cairns T, Goldman G, Armstrong-James D, Haynes K, Haas H, Schrettl M, May G, Nierman WC, Bignell E (2008) Sub-telomere directed gene expression during initiation of invasive aspergillosis. PLoS Pathog 4(9):e1000154

Mendes NS, Silva PM, Silva-Rocha R, Martinez-Rossi NM, Rossi A (2016) Pre-mRNA splicing is modulated by antifungal drugs in the filamentous fungus *Neurospora crassa*. FEBS Open Bio 6(4):358–368

Mendes NS, Bitencourt TA, Sanches PR, Silva-Rocha R, Martinez-Rossi NM, Rossi A (2018) Transcriptome-wide survey of gene expression changes and alternative splicing in *Trichophyton rubrum* in response to undecanoic acid. Sci Rep 8(1):2520

Missall TA, Lodge JK, McEwen JE (2004) Mechanisms of resistance to oxidative and nitrosative stress: implications for fungal survival in mammalian hosts. Eukaryot Cell 3(4):835–846

Mittal J, Ponce MG, Gendlina I, Nosanchuk JD (2019) *Histoplasma capsulatum*: mechanisms for pathogenesis. Curr Top Microbiol Immunol 422:157–191

Monod M (2008) Secreted proteases from dermatophytes. Mycopathologia 166(5–6):285–294

Morschhauser J, Barker KS, Liu TT, Bla BWJ, Homayouni R, Rogers PD (2007) The transcription factor Mrr1p controls expression of the MDR1 efflux pump and mediates multidrug resistance in *Candida albicans*. PLoS Pathog 3(11):e164

Morton CO, Varga JJ, Hornbach A, Mezger M, Sennefelder H, Kneitz S, Kurzai O, Krappmann S, Einsele H, Nierman WC, Rogers TR, Loeffler J (2011) The temporal dynamics of differential gene expression in *Aspergillus fumigatus* interacting with human immature dendritic cells in vitro. PLoS One 6(1):e16016

Muller V, Viemann D, Schmidt M, Endres N, Ludwig S, Leverkus M, Roth J, Goebeler M (2007) *Candida albicans* triggers activation of distinct signaling pathways to establish a proinflammatory gene expression program in primary human endothelial cells. J Immunol 179(12):8435–8445

Mullick A, Elias M, Harakidas P, Marcil A, Whiteway M, Ge B, Hudson TJ, Caron AW, Bourget L, Picard S, Jovcevski O, Massie B, Thomas DY (2004) Gene expression in HL60 granulocytoids and human polymorphonuclear leukocytes exposed to *Candida albicans*. Infect Immun 72(1):414–429

Munoz JF, Gade L, Chow NA, Loparev VN, Juieng P, Berkow EL, Farrer RA, Litvintseva AP, Cuomo CA (2018) Genomic insights into multidrug-resistance, mating and virulence in *Candida auris* and related emerging species. Nat Commun 9(1):5346

Munoz JF, Delorey T, Ford CB, Li BY, Thompson DA, Rao RP, Cuomo CA (2019) Coordinated host-pathogen transcriptional dynamics revealed using sorted subpopulations and single macrophages infected with *Candida albicans*. Nat Commun 10(1):1607

Neves-da-Rocha J, Bitencourt TA, Oliveira VM, Sanches PR, Rossi A, Martinez-Rossi NM (2019) Alternative splicing in heat shock protein transcripts as a mechanism of cell adaptation in *Trichophyton rubrum*. Cell 8(10)

Nobile CJ, Fox EP, Nett JE, Sorrells TR, Mitrovich QM, Hernday AD, Tuch BB, Andes DR, Johnson AD (2012) A recently evolved transcriptional network controls biofilm development in *Candida albicans*. Cell 148(1–2):126–138

O'Meara TR, Xu W, Selvig KM, O'Meara MJ, Mitchell AP, Alspaugh JA (2013) The *Cryptococcus neoformans* Rim101 transcription factor directly regulates genes required for adaptation to the host. Mol Cell Biol 34(4):673–684

Oosthuizen JL, Gomez P, Ruan J, Hackett TL, Moore MM, Knight DA, Tebbutt SJ (2011) Dual organism transcriptomics of airway epithelial cells interacting with conidia of *Aspergillus fumigatus*. PLoS One 6(5):e20527

Osborne CS, Leitner I, Favre B, Ryder NS (2005) Amino acid substitution in *Trichophyton rubrum* squalene epoxidase associated with resistance to terbinafine. Antimicrob Agents Chemother 49(7):2840–2844

Osborne CS, Leitner I, Hofbauer B, Fielding CA, Favre B, Ryder NS (2006) Biological, biochemical, and molecular characterization of a new clinical *Trichophyton rubrum* isolate resistant to terbinafine. Antimicrob Agents Chemother 50(6):2234–2236

Paiao FG, Segato F, Cursino-Santos JR, Peres NT, Martinez-Rossi NM (2007) Analysis of *Trichophyton rubrum* gene expression in response to cytotoxic drugs. FEMS Microbiol Lett 271(2):180–186

Pathakumari B, Liang G, Liu W (2020) Immune defence to invasive fungal infections: a comprehensive review. Biomed Pharmacother 130:110550

Peres NTA, Maranhao FCA, Rossi A, Martinez-Rossi NM (2010a) Dermatophytes: host-pathogen interaction and antifungal resistance. An Bras Dermatol 85(5):657–667

Peres NTA, Sanches PR, Falcão JP, Silveira HCS, Paião FG, Maranhão FCA, Gras DE, Segato F, Cazzaniga RA, Mazucato M, Cursino-Santos JR, Aquino-Ferreira R, Rossi A, Martinez-Rossi NM (2010b) Transcriptional profiling reveals the expression of novel genes in response to various stimuli in the human dermatophyte *Trichophyton rubrum*. BMC Microbiol 10:39–48

Peres NT, Silva LG, Santos Rda S, Jacob TR, Persinoti GF, Rocha LB, Falcao JP, Rossi A, Martinez-Rossi NM (2016) *In vitro* and *ex vivo* infection models help assess the molecular aspects of the interaction of *Trichophyton rubrum* with the host milieu. Med Mycol 54(4):420–427

Perlin DS (2015) Mechanisms of echinocandin antifungal drug resistance. Ann N Y Acad Sci 1354:1–11

Persinoti GF, de Aguiar Peres NT, Jacob TR, Rossi A, Vencio RZ, Martinez-Rossi NM (2014) RNA-sequencing analysis of *Trichophyton rubrum* transcriptome in response to sublethal doses of acriflavine. BMC Genomics 15(Suppl 7):S1

Petrucelli MF, Peronni K, Sanches PR, Komoto TT, Matsuda JB, Silva Junior WAD, Beleboni RO, Martinez-Rossi NM, Marins M, Fachin AL (2018) Dual RNA-Seq analysis of *Trichophyton rubrum* and HaCat keratinocyte co-culture highlights important genes for fungal-host interaction. Genes (Basel) 9(7)

Petrucelli MF, Matsuda JB, Peroni K, Sanches PR, Silva WA Jr, Beleboni RO, Martinez-Rossi NM, Marins M, Fachin AL (2019) The transcriptional profile of *Trichophyton rubrum* co-cultured with human keratinocytes shows new insights about gene modulation by terbinafine. Pathogens 8(4)

Ramirez MA, Lorenz MC (2007) Mutations in alternative carbon utilization pathways in *Candida albicans* attenuate virulence and confer pleiotropic phenotypes. Eukaryot Cell 6(2):280–290

Richie DL, Hartl L, Aimanianda V, Winters MS, Fuller KK, Miley MD, White S, McCarthy JW, Latge JP, Feldmesser M, Rhodes JC, Askew DS (2009) A role for the unfolded protein response (UPR) in virulence and antifungal susceptibility in *Aspergillus fumigatus*. PLoS Pathog 5(1):e1000258

Rocha EM, Gardiner RE, Park S, Martinez-Rossi NM, Perlin DS (2006) A Phe389Leu substitution in ergA confers terbinafine resistance in *Aspergillus fumigatus*. Antimicrob Agents Chemother 50(7):2533–2536

Rodrigues ML, Nosanchuk JD (2020) Fungal diseases as neglected pathogens: a wake-up call to public health officials. PLoS Negl Trop Dis 14(2):e0007964

Romani L (2011) Immunity to fungal infections. Nat Rev Immunol 11(4):275–288

Rosam K, Monk BC, Lackner M (2020) Sterol 14alpha-demethylase ligand-binding pocket-mediated acquired and intrinsic azole resistance in fungal pathogens. J Fungi (Basel) 7(1)

Rossi A, Cruz AHS, Santos RS, Silva PM, Silva EM, Mendes NS, Martinez-Rossi NM (2013) Ambient pH sensing in filamentous fungi: pitfalls in elucidating regulatory hierarchical signaling networks. IUBMB Life 65(11):930–935

Rossi A, Martins MP, Bitencourt TA, Peres NTA, Rocha CHL, Rocha FMG, Neves-da-Rocha J, Lopes MER, Sanches PR, Bortolossi JC, Martinez-Rossi NM (2021) Reassessing the use of undecanoic acid as a therapeutic strategy for treating fungal infections. Mycopathologia

Rude TH, Toffaletti DL, Cox GM, Perfect JR (2002) Relationship of the glyoxylate pathway to the pathogenesis of *Cryptococcus neoformans*. Infect Immun 70(10):5684–5694

Sagatova AA (2021) Strategies to better target fungal squalene monooxygenase. J Fungi (Basel) 7(1)

Sanguinetti M, Posteraro B, Fiori B, Ranno S, Torelli R, Fadda G (2005) Mechanisms of azole resistance in clinical isolates of *Candida glabrata* collected during a hospital survey of antifungal resistance. Antimicrob Agents Chemother 49(2):668–679

Santos HL, Lang EAS, Segato F, Rossi A, Martinez-Rossi NM (2018) Terbinafine resistance conferred by multiple copies of the salicylate 1-monooxygenase gene in *Trichophyton rubrum*. Med Mycol 56(3):378–381

Schobel F, Ibrahim-Granet O, Ave P, Latge JP, Brakhage AA, Brock M (2007) *Aspergillus fumigatus* does not require fatty acid metabolism via isocitrate lyase for development of invasive aspergillosis. Infect Immun 75(3):1237–1244

Seelbinder B, Wallstabe J, Marischen L, Weiss E, Wurster S, Page L, Loffler C, Bussemer L, Schmitt AL, Wolf T, Linde J, Cicin-Sain L, Becker J, Kalinke U, Vogel J, Panagiotou G, Einsele H, Westermann AJ, Schauble S, Loeffler J (2020) Triple RNA-Seq reveals synergy in a human virus-fungus co-infection model. Cell Rep 33(7):108389

Segato F, Nozawa SR, Rossi A, Martinez-Rossi NM (2008) Over-expression of genes coding for proline oxidase, riboflavin kinase, cytochrome c oxidase and an MFS transporter induced by acriflavin in *Trichophyton rubrum*. Med Mycol 46(2):135–139

Shapiro RS, Robbins N, Cowen LE (2011) Regulatory circuitry governing fungal development, drug resistance, and disease. Microbiol Mol Biol Rev 75(2):213–267

Shiraki Y, Ishibashi Y, Hiruma M, Nishikawa A, Ikeda S (2006) Cytokine secretion profiles of human keratinocytes during *Trichophyton tonsurans* and *Arthroderma benhamiae* infections. J Med Microbiol 55(Pt 9):1175–1185

Silva SS, Tavares AHFP, Passos-Silva DG, Fachin AL, Teixeira SMR, Soares CMA, Carvalho MJA, Bocca AL, Silva-Pereira I, Passos GAS, Felipe MSS (2008) Transcriptional response of murine macrophages upon infection with opsonized *Paracoccidioides brasiliensis* yeast cells. Microbes Infect 10(1):12–20

Silva DL, Lima CM, Magalhaes VCR, Baltazar LM, Peres NTA, Caligiorne RB, Moura AS, Fereguetti T, Martins JC, Rabelo LF, Abrahao JS, Lyon AC, Johann S, Santos DA (2021) Fungal and bacterial coinfections increase mortality of severely ill COVID-19 patients. J Hosp Infect

Silva MG, Schrank A, Bailão EFLC, Bailão AM, Borges CL, Staats CC, Parente JA, Pereira M, Salem-Izacc SM, Mendes-Giannini MJS, Oliveira RMZ, Rosa e Silva LK, Nosanchuk JD, Vainstein MH and Soares CMA (2011) The homeostasis of iron, copper, and zinc in *Paracoccidioides brasiliensis, Cryptococcus neoformans* var. *grubii*, and *Cryptococcus gattii*: a comparative analysis. Front Microbiol 2:49. https://doi.org/10.3389/fmicb.2011.00049

Silveira HCS, Gras DE, Cazzaniga RA, Sanches PR, Rossi A, Martinez-Rossi NM (2010) Transcriptional profiling reveals genes in the human pathogen *Trichophyton rubrum* that are expressed in response to pH signaling. Microb Pathog 48(2):91–96

Souto JT, Figueiredo F, Furlanetto A, Pfeffer K, Rossi MA, Silva JS (2000) Interferon-gamma and tumor necrosis factor-alpha determine resistance to *Paracoccidioides brasiliensis* infection in mice. Am J Pathol 156(5):1811–1820

Staib P, Zaugg C, Mignon B, Weber J, Grumbt M, Pradervand S, Harshman K, Monod M (2010) Differential gene expression in the pathogenic dermatophyte *Arthroderma benhamiae* in vitro versus during infection. Microbiology 156(Pt 3):884–895

Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: the teenage years. Nat Rev Genet 20(11):631–656

Subramani A, Griggs P, Frantzen N, Mendez J, Tucker J, Murriel J, Sircy LM, Millican GE, McClelland EE, Seipelt-Thiemann RL, Nelson DE (2020) Intracellular *Cryptococcus neoformans* disrupts the transcriptome profile of M1- and M2-polarized host macrophages. PLoS One 15(8):e0233818

Sugui JA, Kim HS, Zarember KA, Chang YC, Gallin JI, Nierman WC, Kwon-Chung KJ (2008) Genes differentially expressed in conidia and hyphae of *Aspergillus fumigatus* upon exposure to human neutrophils. PLoS One 3(7):e2655

Sun N, Fonzi W, Chen H, She X, Zhang L, Calderone R (2013) Azole susceptibility and transcriptome profiling in *Candida albicans* mitochondrial electron transport chain complex I mutants. Antimicrob Agents Chemother 57(1):532–542

Tavares AH, Derengowski LS, Ferreira KS, Silva SS, Macedo C, Bocca AL, Passos GA, Almeida SR, Silva-Pereira I (2012) Murine dendritic cells transcriptional modulation upon *Paracoccidioides brasiliensis* infection. PLoS Negl Trop Dis 6(1):e1459

Thewes S, Kretschmar M, Park H, Schaller M, Filler SG, Hube B (2007) In vivo and ex vivo comparative transcriptional profiling of invasive and non-invasive *Candida albicans* isolates identifies genes associated with tissue invasion. Mol Microbiol 63(6):1606–1628

Tierney L, Linde J, Muller S, Brunke S, Molina JC, Hube B, Schock U, Guthke R, Kuchler K (2012) An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. Front Microbiol 3:85

Upadhya R, Kim H, Jung KW, Park G, Lam W, Lodge JK, Bahn YS (2013) Sulphiredoxin plays peroxiredoxin-dependent and -independent roles via the HOG signalling pathway in *Cryptococcus neoformans* and contributes to fungal virulence. Mol Microbiol 90(3):630–648

Vediyappan G, Rossignol T, d'Enfert C (2010) Interaction of *Candida albicans* biofilms with antifungals: transcriptional response and binding of antifungals to beta-glucans. Antimicrob Agents Chemother 54(5):2096–2111

Vermes A, Guchelaar HJ, Dankert J (2000) Flucytosine: a review of its pharmacology, clinical indications, pharmacokinetics, toxicity and drug interactions. J Antimicrob Chemother 46(2):171–179

Wachtler B, Wilson D, Haedicke K, Dalle F, Hube B (2011) From attachment to damage: defined genes of *Candida albicans* mediate adhesion, invasion and damage during interaction with oral epithelial cells. PLoS One 6(2):e17046

Walker LA, Maccallum DM, Bertram G, Gow NA, Odds FC, Brown AJ (2009) Genome-wide analysis of *Candida albicans* gene expression patterns during infection of the mammalian kidney. Fungal Genet Biol 46(2):210–219

Wang M, Zhao Y, Cao L, Luo S, Ni B, Zhang Y, Chen Z (2021) Transcriptome sequencing revealed the inhibitory mechanism of ketoconazole on clinical *Microsporum canis*. J Vet Sci 22(1):e4

Westermann AJ, Barquist L, Vogel J (2017) Resolving host-pathogen interactions by dual RNA-seq. PLoS Pathog 13(2):e1006033

White TC, Holleman S, Dy F, Mirels LF, Stevens DA (2002) Resistance mechanisms in clinical isolates of *Candida albicans*. Antimicrob Agents Chemother 46(6):1704–1713

Woodfolk JA, Platts-Mills TA (1998) The immune response to dermatophytes. Res Immunol 149(4–5):436–445; discussion 522-433

Yamada T, Maeda M, Alshahni MM, Tanaka R, Yaguchi T, Bontems O, Salamin K, Fratti M, Monod M (2017) Terbinafine resistance of *Trichophyton* clinical isolates caused by specific point mutations in the squalene epoxidase gene. Antimicrob Agents Chemother 61(7)

Yu L, Zhang W, Liu T, Wang X, Peng J, Li S, Jin Q (2007a) Global gene expression of *Trichophyton rubrum* in response to PH11B, a novel fatty acid synthase inhibitor. J Appl Microbiol 103(6):2346–2352

Yu L, Zhang W, Wang L, Yang J, Liu T, Peng J, Leng W, Chen L, Li R, Jin Q (2007b) Transcriptional profiles of the response to ketoconazole and amphotericin B in *Trichophyton rubrum*. Antimicrob Agents Chemother 51(1):144–153

Yu CH, Chen Y, Desjardins CA, Tenor JL, Toffaletti DL, Giamberardino C, Litvintseva A, Perfect JR, Cuomo CA (2020) Landscape of gene expression variation of natural isolates of *Cryptococcus neoformans* in response to biologically relevant stresses. Microb Genom 6(1)

Zakikhany K, Naglik JR, Schmidt-Westhausen A, Holland G, Schaller M, Hube B (2007) In vivo transcript profiling of *Candida albicans* identifies a gene essential for interepithelial dissemination. Cell Microbiol 9(12):2938–2954

Zamith-Miranda D, Amatuzzi RF, Martins ST, Vieira AZ, da Rocha IM, Rodrigues ML, Trentin G, Almeida F, Nakayasu ES, Nosanchuk JD, Alves LR (2020) Integrated transcriptional analysis of the cellular and extracellular vesicle RNA content of *Candida auris* in response to caspofungin. bioRxiv. https://doi.org/10.1101/2020.12.04.411843

Zhang W, Yu L, Yang J, Wang L, Peng J, Jin Q (2009) Transcriptional profiles of response to terbinafine in *Trichophyton rubrum*. Appl Microbiol Biotechnol 82(6):1123–1130

Zhu X, Ge Y, Wu T, Zhao K, Chen Y, Wu B, Zhu F, Zhu B, Cui L (2020) Co-infection with respiratory pathogens among COVID-2019 cases. Virus Res 285:198005

Chapter 18
Understanding Chagas Disease by Multi-omics Data Integration, Functional, and Enrichment Computational Analysis

Ludmila Rodrigues Pinto Ferreira

18.1 Epidemiology of Chagas Disease

Chagas disease, also called human American trypanosomiasis, was *named* after the Brazilian medical doctor Carlos Chagas (Fig. 18.1a), who discovered the disease in 1909 during a campaign to fight malaria in Brazil (Moncayo 2010).

Carlos Chagas identified, associated with diseased individuals living in poor dwellings (Fig. 18.1b), a triatomine blood-sucking insect (Fig. 18.1c). He found flagellated parasites in the intestine of the bug, which he named *Trypanosoma cruzi* (*T. cruzi*) (Fig. 18.1d). He also found *T. cruzi* parasites in the blood of sick people, and soon correlated the parasitemia (level of parasites in the blood) with some symptoms of the disease, such as fever, anemia, lymphadenopathy, splenomegaly, and a cardiac form of the disease (Kropf and Sa 2009; Pays 2009; Kropf 2011).

The disease begins with a short acute phase characterized by high parasitemia followed by a life-long chronic phase maintained with scarce parasites (Golgher and Gazzinelli 2004). The World Health Organization (WHO) estimated that 8 million people are infected worldwide, mostly in Latin America.

Over 25 million people are at risk of the disease, and 10000 people die every year from clinical manifestations of Chagas disease (Hotez et al. 2012). Natural transmission of Chagas disease has been controlled in many countries by insecticide targeting of hematophagous bug populations, as well as improved socioeconomic status and quality of dwelling in Latin America.

The list of possible infection routes of Chagas disease includes vectorial, transfusional (through *T. cruzi* infected blood), congenital, through organ transplantation, oral transmission, and accidental, through laboratory accidents. In 2006, WHO

L. R. P. Ferreira (✉)
RNA Systems Biology Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brazil
e-mail: ludmila@icb.ufmg.br

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_18

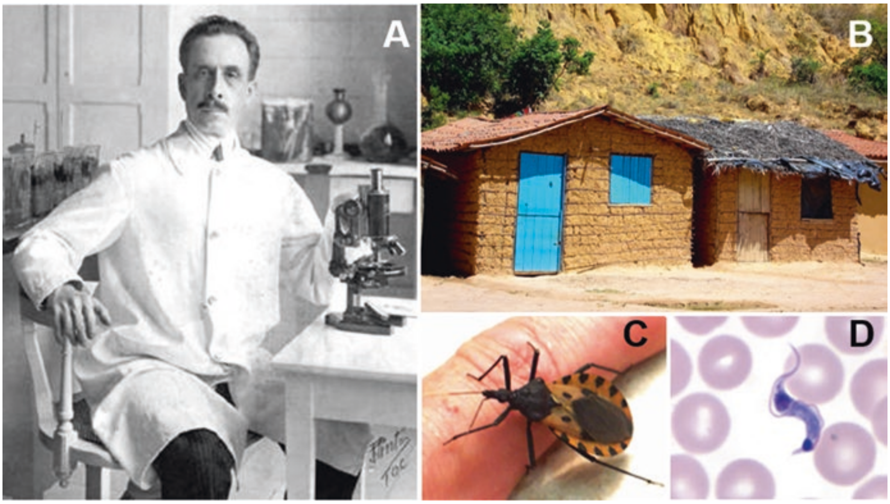


Fig. 18.1 Chagas disease, also called human American trypanosomiasis, was n*amed* after the Brazilian medical doctor Carlos Chagas (**a**). He identified, associated with diseased individuals living in poor dwellings (**b**), a triatomine blood-sucking insect (**c**). He found flagellated parasites in the intestine of the bug, which he named *Trypanosoma cruzi* (*T. cruzi*; trypomastigote in a thin blood smear stained with Giemsa (**d**). (Image credits: (**a**) Public domain, (**b**) José Eduardo R. Camargo (**c**) and (**d**) Public Health Image Library – Centers for Disease Control and Prevention – CDC and Laboratory Identification of Parasitic Diseases of Public Health Concern – DPDx)

certified Brazil as being free of transmission through *Triatoma infestans*, the main intradomicilliary vector of Chagas disease (WHO Expert Committee 2002).

However, there were new reports of oral transmission and oral outbreaks in the Amazon region showing that this victory was only partial (Dias 2009). Despite improvement in quality assurance of blood transfusions and organ transplants, Chagas disease remains a public health problem in Latin America and is becoming a new threat for *T. cruzi* infection in non-endemic countries (Diaz 2007; Lescure et al. 2010; Hotez et al. 2012).

It has been estimated that 700,000 infected people are living outside of Latin America (Hotez et al. 2012). The pathogenesis of Chagas disease remains largely unknown, and there are still no effective vaccines or drugs to prevent or treat chronic infection with *T. cruzi*.

18.2 *T. cruzi* Life Cycle and Triatomine Vectors

T. cruzi is known to infect eight different mammalian orders including humans, and it is transmitted by insect vectors of the Reduviidae family and the subfamily of Triatomines (Committee 2002). There are many popular names for the vector. In Brazil the common name for the vector is *Barbeiro* – “the barber” and the English name is the “kissing bug.”

Around 100 different triatomine species are susceptible to infection with the *T. cruzi* parasite but the principal vector species has been *Triatoma infestans* and in Brazil, the species *Triatoma sordida* and *Panstrongylus megistus* are also prevalent (Moncayo and Ortiz Yanine 2006; Dias 2009; Siqueira-Batista et al. 2011). *T. cruzi* has different developmental stages in its life cycle: Epimastigotes are the form stage that proliferates by cell division in the stomach of the triatomine bugs, they migrate to the distal part of the bug's intestine, and by a process called metacyclogenesis, they transform into metacyclic trypomastigotes, the infective form for the vertebrate host.

The insects feed on mammals by sucking blood, and *T. cruzi* is transferred via their feces, deposited on the skin of the host after feeding. The metacyclic trypomastigotes can penetrate through mucous membranes as well as skin injuries when the host scratches the skin after being bitten or rub their eye.

The parasites then invade host cells, transforming into amastigotes which replicate and differentiate into trypomastigotes, disrupting host cells and infecting various cell types with a particular tropism for cardiac, skeletal, and smooth muscle cells (de Souza et al. 2010). Finally, the bugs are infected by ingesting trypomastigotes in the blood from infected hosts, thus completing the *T. cruzi* life cycle (Fig. 18.2).

18.3 Basic Knowledge About the Clinical Features of Chagas Disease: Acute and Chronic Phases of Infection

The parasite *T. cruzi* produces pathological processes in mammals that can occur in various organs and tissues. When *T. cruzi* is transmitted, it invades the victim's bloodstream and the lymphatic system. Hereafter, it nestles in many tissues including the skeletal muscle and cardiac tissue, which causes immune responses and inflammation (WHO Expert Committee 2002).

Chagas disease has an acute as well as a chronic phase. Morbidity and mortality are higher in the acute phase for children under five, immunosuppressed people, or people with high parasitemia as in patients from outbreaks of food-borne Chagas disease.

The acute phase can occur at any age in disease endemic areas; however, the highest frequency is before the age of 15, typically starting in the age group 1–5 years. The acute phase of Chagas disease usually lasts 6–8 weeks, and most frequently is oligo- or asymptomatic and after this phase, most patients appear to be healthy (Moncayo and Ortiz Yanine 2006).

The infection by *T. cruzi* can then only be detected by serological or parasitological tests. In the acute phase, if the transmission is vectorial, visible port of entry can be identified, such as the chagoma, a skin lesion in exposed areas of the body, or the Romaña's sign, a purplish edema on the lids of one eye (Fig. 18.3). The sign occurs only in about 10% of infected persons, and can easily be misdiagnosed with

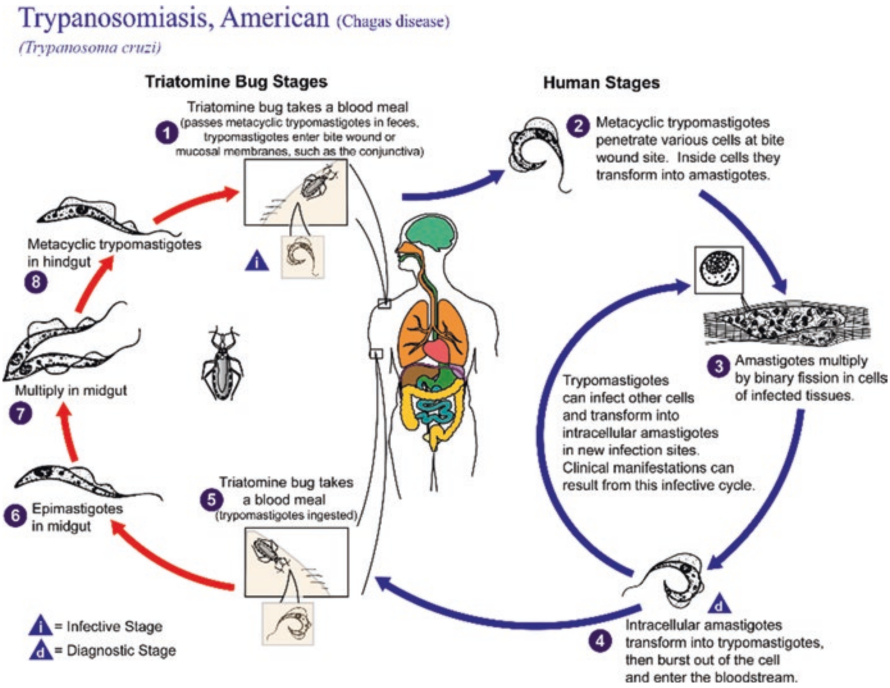

Trypanosomiasis, American (Chagas disease)
(Trypanosoma cruzi)
Triatomine Bug Stages
Human Stages
1 Triatomine bug takes a blood meal (passes metacyclic trypomastigotes in feces, trypomastigotes enter bite wound or mucosal membranes, such as the conjunctiva)
2 Metacyclic trypomastigotes penetrate various cells at bite wound site. Inside cells they transform into amastigotes.
3 Amastigotes multiply by binary fission in cells of infected tissues.
4 Intracellular amastigotes transform into trypomastigotes, then burst out of the cell and enter the bloodstream.
5 Triatomine bug takes a blood meal (trypomastigotes ingested)
6 Epimastigotes in midgut
7 Multiply in midgut
8 Metacyclic trypomastigotes in hindgut
Trypomastigotes can infect other cells and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle.
i
d
i = Infective Stage
d = Diagnostic Stage

Fig. 18.2 *T. cruzi* life cycle: An infected triatomine insect vector takes a blood meal and releases trypomastigotes in its feces near the site of the bite wound. Trypomastigotes enter the host through the wound or through intact mucosal membranes, such as the conjunctiva. Inside the host, the trypomastigotes invade cells near the site of inoculation, where they differentiate into intracellular amastigotes. The amastigotes multiply by binary fission and differentiate into trypomastigotes, and then are released into the circulation as bloodstream trypomastigotes. The "kissing" bug becomes infected by feeding on human or animal blood that contains circulating parasites. The ingested trypomastigotes transform into epimastigotes in the vector's midgut. The parasites multiply and differentiate in the midgut and differentiate into infective metacyclic trypomastigotes in the hindgut. (Life cycle image and information credit: Laboratory Identification of Parasitic Diseases of Public Health Concern – DPDx (http://www.cdc.gov/dpdx))

conjunctivitis, for example, which is common in rural areas (Roveda 1967; Delaporte 1997; Dias 1997).

Other clinical features of the acute phase are an excessive activation of the immune system that includes cytokinemia (high plasma levels of cytokines), intense activation of B and T cells. Generic and unspecific symptoms include diarrhea, vomiting, headache, muscle pain, loss of appetite, and extreme fatigue.

These symptoms are not very specific and can easily be confused with other disease etiologies (Coura and Borges-Pereira 2010). Shortly after the acute infection starts, *T. cruzi* components – including its DNA and membrane glycoconjugates – trigger innate immunity via Toll-like receptors in macrophages and dendritic cells, among other cell types. Upon activation, cells from monocytic lineage

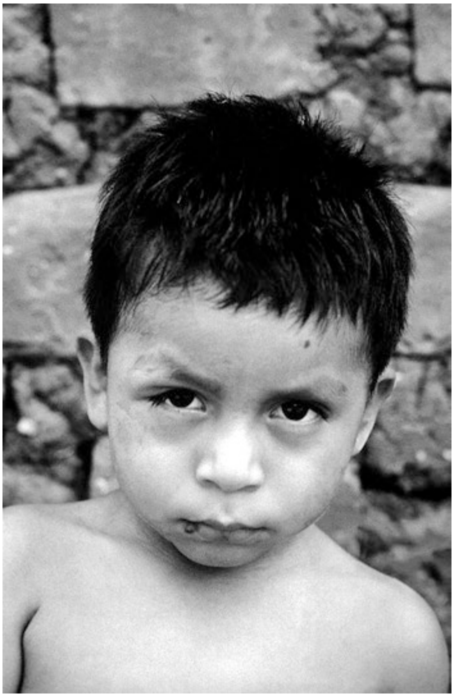


Fig. 18.3 Romaña's sign, a purplish edema on the lids of one eye that is formed during T. cruzi infection. (The illustrations of chagasic patient were obtained from: Public Health Image Library – Centers for Disease Control and Prevention – CDC/Dr. Mae Melvin)

produce high levels of proinflammatory cytokines like interferon gamma (IFN-γ), interleukin 12 (IL-12), and tumor necrosis factor alpha (TNF-α).

The high level of IFN-γ-induced chemokines and adhesion molecules plays an important role in promoting the inflammatory environment in the heart of animals infected with *T. cruzi*.

In fact, mice lacking the functional IFN-γ gene display major changes in the CD4+ T and CD8+ T lymphocytes composition of inflammatory infiltrates, as well as enhanced tissue parasitism in the heart (Campos et al. 2004). The essential role of some of these cytokines (e.g., IL-12 and TNF-α) and reactive nitrogen intermediates (RNI) in the control of parasitemia and tissue parasitism is evidenced during the early stages of infection in the murine model (Junqueira et al. 2010).

More precisely, the cells from the macrophage lineage exposed to *T. cruzi* will produce IL-12 that is responsible for initiating IFN-γ synthesis by natural killer (NK) cell. IFN-γ plays a major role in resistance through the activation of macrophages to produce high levels of RNI that will effectively control parasite replication (Fig. 18.4).

If not controlled by the innate immune system of the host, the infection is fatal as shown in experimental models employing mice lacking functional genes for the IL-12, IFN-γ, IFN-γ receptor, TNF-α receptor, or inducible nitric oxide (NO) synthase (iNOS) genes (Golgher and Gazzinelli 2004; Gazzinelli and Denkers 2006; Junqueira et al. 2010).

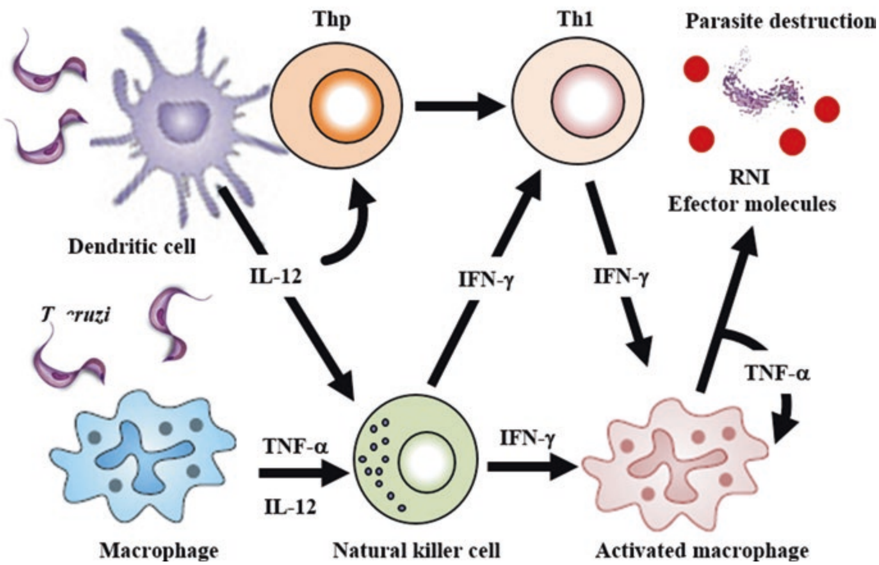


Fig. 18.4 Immune response to *T. cruzi* infection. In the initial stage of *T. cruzi* invasion, cells from the innate immune system [dendritic cell, macrophages, and natural killer cells (NK cells)] produce cytokines (IL-12, TNF-α, and IFN-γ) and effector molecules [reactive nitrogen intermediates (RNIs)] that lead to parasite destruction. At the same time, innate immune cells, particularly dendritic cells, make the bridge between the innate and acquired immunity, producing cytokines (IL-12) necessary for differentiation and clonal expansion of T helper 1 (Th1) CD4+. IFN-γ produced by CD4+ activates effector mechanisms in macrophages to destroy both amastigotes and phagocytosed trypomastigotes. Abbreviations: *IFN* interferon, *IL* interleukin, *Thp* Th precursor cell, *TNF* tumor necrosis factor

The chronic phase starts with an effective acquired immunity leading to parasitemia drop to a level where it is undetectable with direct parasitological tests, and when symptoms and clinical manifestations typically disappear.

However, depending on different factors 10–40% of patients in the chronic phase will develop lesions in target organs, like the intestine (intestinal mega syndrome), esophagus (mega esophagus), and heart (cardiomyopathy); however, up to 70% of infected people remain in an indeterminate asymptomatic form (ASY) for their whole life. The most important clinical consequence of chronic Chagas disease is the chronic Chagas disease cardiomyopathy (CCC), an inflammatory cardiomyopathy that develops in up to 30% of infected individuals.

A significant proportion of those patients subsequently develop dilated cardiomyopathy with a fatal outcome. Heart failure of chagasic etiology has a worse prognosis and 50% lower survival rate than cardiomyopathies of noninflammatory etiology, like ischemic (IC) and idiopathic dilated cardiomyopathy (DCM) (Bilate and Cunha-Neto 2008; Machado et al. 2012).

Inflammatory cytokines are produced during the chronic phase of Chagas disease. Mononuclear cells increase their cytokine production, leading to increased plasma levels of TNF-α and IFN-γ, and are even detected in infected ASY individuals, probably in response to parasite persistence.

The subset of patients that develop CCC displays an array of immunological alterations consistent with an exacerbated Th1 immune response; the predominance of production of IFN-γ and TNF-α (Abel et al. 2001) associated to the increased expression of the Th1 transcription factor T-bet in the heart, which is not controlled by regulatory T cells in situ is evidence corroborating that the Th1 response is involved in tissue damage in CCC.

Chagas thus remains a neglected disease, with no vaccines or antiparasitic drugs proven efficient in chronically infected adults, when most patients are diagnosed. Development of effective drugs for CCC is hampered by the limited knowledge of its pathogenesis. T cell migration to the myocardium and inflammation, cytokine/chemokine-induced modulation of myocardial gene and protein expression, and genetic components controlling such processes are clearly key events (Nogueira et al. 2012).

Regardless of the mechanisms underlying the initiation and maintenance of the myocarditis, the bulk of the evidence indicates that the inflammatory infiltrate is a significant effector of heart tissue damage (Coura and Borges-Pereira 2010).

18.4 High-Throughput Analysis Helping to Understand the Mechanisms Involved in Chagas Disease

The pathogenesis of CCC is still matter of intense debate. The susceptibility factors that lead to 30% of individuals to develop CCC after *T. cruzi* infection remain unknown, and it is a challenge to identify patients who are at risk of dying. The absence of an alternative treatment for CCC demonstrates that our knowledge about its pathogenesis is still very limited and that new study strategies are needed to discover biomarkers of disease progression as well as new treatments for CCC.

The first gene expression analyses in Chagas disease were performed primarily based on observations from immunoblotting, polymerase chain reaction, and/or northern blotting limited to evaluation of a few preselected genes at one time (Ferreira et al. 1999; Ferreira et al. 2002).

Another limitation was the access to human heart samples from the acute phase of the disease, so most data available is based on murine models and/or using cells from in vitro *T. cruzi* infection. Several reports have been published with these approaches with a high variability in parasite strains, host cells, mammalian species, and times of infection generating a complex picture and few general conclusions.

The continuous advance of transcriptome analysis techniques, from different types of microarrays to RNA sequencing (RNAseq) and other omics techniques, expanded dramatically in the past few years revolutionized the field of molecular biology and afforded the opportunity to profile the expression of thousands of genes, collecting a large amount of data, permitting the identification of new molecular players in the pathogenesis of Chagas disease.

Next, we will describe some of the important studies on Chagas disease over the years, based on analysis of gene expression and also the recent findings describing microRNAs as new players in the disease pathogenesis and also adding complexity in the biological system demanding computational and systems biology approaches to understand and translate the large amount of data produced in knowledge.

18.5 Transcriptome Analysis in In Vitro Models

In 2002, Burleigh et al. have performed the first microarray analysis to identify differences in gene expression using an in vitro model of human fibroblasts infected with *T. cruzi* (Vaena de Avalos et al. 2002). For this experiment, they used a glass slides high-density microarray consisting of ~27,000 human cDNAs that were hybridized with fluorescent probes generated from *T. cruzi*-infected human fibroblasts (HFF) at early time points following infection (2–24 h). Surprisingly, they observed that no genes were induced ≥2-fold in HFF cDNA between 2 and 6 h postinfection (hpi).

A significant increase in transcript abundance for 106 host cell genes was observed only at 24 hpi. Among the most highly induced was a set of interferon-stimulated genes, indicative of a type I interferon (IFN) response to *T. cruzi*. The authors concluded that the delay of *T. cruzi* to induce host cell transcriptional responses is indicative that changes in host cell gene expression may correlate with a particular parasite-dependent event such as differentiation or replication. These events are performed by *T. cruzi* silently without eliciting major changes in the host fibroblasts gene expression.

Because cardiac myocytes are important targets of initial infection with *T. cruzi*, in 2009 another study compared gene profiling of primary cultures of cardiac myocytes infected for 48 hours with *T. cruzi* (Goldenberg et al. 2009). They employed microarray analysis with glass slides containing a total of 31,769 70mer oligonucleotide probes. As expected, the results are diverse from the study done using fibroblasts and show a substantial alteration in expression of more than 5% of the sampled genome with major alterations in genes related to inflammation, immunological responses, and cell adhesion.

Among the pathways most affected from the list of upregulated genes were those involved in enzymatic activity, immune and stress responses, apoptosis and activation of the proteasome, and calcium-activated potassium channel activity. Downregulated pathways included calcium and second messenger signaling, cytoskeleton elements (actin filaments, stress fibers, myosin), enzymatic degradation (lysozyme, trypsin, metallopeptidases), and extracellular matrix. This study showed that the cardiac myocytes themselves contribute to the remodeling process even in the absence of other confounding factors, even though in vivo models show contributions by fibroblasts and heart-infiltrating inflammatory cells.

18.6 Transcriptome and Proteome Analysis in Rodent Models

The study of chagasic heart disease has been aided using the mouse model which recapitulates many of the functional and pathological alterations of the human disease. In 2003, two different groups have performed microarray analysis to detect differences in gene expression in the heart of mice experimentally infected with *T. cruzi*. They have used different microarray platforms, i.e., Garg et al. have used commercial nylon membranes microarrays containing a repertoire of 1,176 mouse genes printed on the arrays to evaluate the gene expression in whole heart of mice infected with SylvioX10/4 strain of *T. cruzi* for 3, 37, and 110 days postinfection (dpi) (Garg et al. 2003), and Mukherjee et al. have used glass microarrays containing ~27,400 mouse cDNAs clones to evaluate gene expression also in whole heart from C57BL/6 129sv mice infected for 100 days with the Brazil strain of *T. cruzi* (Mukherjee et al. 2003).

Garg et al. showed that out of a total of 1176 genes printed on the arrays, 31, 89, and 66 genes were differentially regulated in the context of their expression trends at 3, 37, and 110 dpi, respectively. They showed that all the differentially expressed genes in the myocardium at 3 dpi were upregulated and encoded immune-related or host defense/stress proteins. During the acute phase (37 dpi), mRNA species for 77 of the 89 differentially regulated genes were increased by at least twofold. Of these, 27 transcripts were increased by >10-fold, and 18 of the 27 transcripts encoded the immune-related proteins. Out of the 12 transcripts that were reproducibly repressed at 37 dpi, eight were characterized to encode proteins involved in mitochondrial energy metabolism.

Surprisingly, a majority of the differentially expressed genes (>63%) in the myocardium of infected mice at 110 dpi were repressed relative to normal controls. From the 66 differentially expressed gene at 110 dpi, 42 were repressed and of these, 26 (60%) transcripts have implications in sustaining the mitochondrial energy metabolism and maintaining the cytoskeletal and extracellular matrix (ECM) structure and function.

The study performed by Mukherjee et al. (2003) also demonstrated the induction of several genes important to cardiac remodeling, like cytokines and growth factor genes, including growth differentiation factor 3 and insulin-like growth factor-binding proteins, a family of structurally homologous secreted proteins that specifically bind and modulate the activities of insulin-like growth factors (IGF-1 and IGF-2), enhance cellular differentiation and stimulate cell proliferation and muscle cell differentiation.

Results from both studies are in accordance showing changes in oxidative phosphorylation and depressed energy metabolism. Soares et al. (2010) also analyzed gene expression profiling in total heart from C57Bl/6 mice chronically infected (8 months of infection) with *T. cruzi* (Colombiana strain) (Soares et al. 2010). They used, for their analysis, glass slides microarrays spotted with 32,620 mouse 70mer oligonucleotides. Their results showed some similarities to the previous studies. As expected, mice chronically infected with *T. cruzi* have intense myocarditis, with an inflammatory infiltrate mainly composed by mononuclear cells, including CD4+ and CD8+ T lymphocytes and macrophages.

So, the arrays showed alterations in a great number of genes related to inflammation and immune responses. Genes coding for the macrophage cell surface marker CD68 and lymphocytes antigens CD38 and CD 52 had their expression increased, a finding compatible with the presence of these cells in the inflammatory infiltrate. The expression of genes coding for adhesion molecules, such as galectin-3, P-selectin ligand (CD162), integrin β3 (CD61), and ICAM-1 (CD54), was increased in hearts of chagasic mice.

Cytokine-associated genes were differentially expressed in hearts of chagasic mice like IFNγ and TNF-α. Another characteristic of hearts in chronically chagasic mice is fibrosis. The results showed upregulation of genes related to synthesis of extracellular matrix components and an increased expression of lysyl oxidase, an enzyme that promotes the cross-linking of collagen fibers.

The tissue inhibitor of metalloproteinase 1 (TIMP-1), an inhibitor of collagen degradation, was also upregulated in chronic chagasic hearts. Bilate et al. have performed a proteomic analysis in hearts of acutely *T. cruzi* infected Syrian hamsters and have shown that severe acute infection is associated to differential expression of structural/contractile and stress response proteins that may be associated with alterations in the cardiomyocyte cytoskeleton (Bilate et al. 2008).

18.7 Transcriptomics in CCC Hearts Explanted During Heart Transplantation

In 2005, Cunha-Neto et al. showed the first gene expression profiling study in human heart samples from Chagas patients and controls, obtained at transplantation (Cunha-Neto et al. 2005). They used a 10,386-element cDNA microarray, built from cardiovascular cDNA libraries, in combination with real-time reverse transcriptase polymerase chain reaction analysis to compare the gene expression fingerprint of five patients with CCC (serological diagnosis, positive epidemiology), seven with DCM (dilated cardiomyopathy in the absence of ischemic disease, and negative epidemiology), and four normal adult heart tissue (obtained from four non-failing donor hearts not used for cardiac transplantation due to size mismatch with available recipients).

They found that gene expression patterns are markedly different in CCC and DCM, with significant activity of IFN-inducible genes in CCC patients. Indeed, it showed that immune response, lipid metabolism, and mitochondrial oxidative phosphorylation genes were selectively up-regulated in myocardial tissue of the tested Chagas' cardiomyopathy patients. Interferon (IFN)-γ-inducible genes represented 15% of genes specifically upregulated in Chagas' cardiomyopathy myocardial tissue, indicating the importance of IFN-γ signaling also in the human model. They also tested whether IFN-γ can directly modulate cardiomyocyte gene expression by exposing fetal murine cardiomyocytes to IFN-γ and the IFN-γ-inducible chemokine monocyte chemoattractant protein-1. Atrial natriuretic factor expression increased 15-fold in response to IFN-γ whereas combined IFN-γ and monocyte chemoattractant protein-1 increased atrial natriuretic factor expression 400-fold. The authors concluded that IFN-γ and chemokine signaling may directly upregulate cardiomyocyte expression of genes involved in pathological hypertrophy, which may lead to heart failure.

Another important result was similar to it was observed in the gene expression analysis in the murine model of *T. cruzi* infection: They saw that IFN-γ and *T. cruzi* infection can depress energy metabolism, thus reducing myocardial ATP generation, which has potential consequences for myocardial contractility, electric conduction, and rhythm.

18.8 MicroRNAs, New Players in Chagas Disease Pathogenesis

Small, noncoding RNAs, known as microRNAs (miRNAs), play a key role in determining which genes are expressed. MiRNAs regulate tissue-specific protein expression and are involved in virtually all cellular processes; up to one-third of mammalian mRNAs are susceptible to miRNA-mediated regulation (Lewis et al. 2003). MiRNAs bind to partially complementary sequences present in the 3' untranslated regions (UTR) of specific "target" mRNA (Agarwal et al. 2015).

This pairing between the miRNA and its target mRNA leads to cleavage of the target mRNA or translation inhibition, resulting in silencing of gene expression. It has been shown that miRNAs are determinants of the physiology and pathophysiology of the cardiovascular system and altered expression of muscle- and/or cardiac-specific miRNAs such as the miRNAs named miR-1, miR-208, and miR-133 in myocardial tissue is involved in heart development and cardiovascular diseases (CD), including myocardial hypertrophy, heart failure, and fibrosis (Chen et al. 2006; Bostjancic et al. 2010; Divakaran 2010; Oliveira-Carvalho et al. 2012).

Several targets of these three miRNAs are related to CD, among them RhoA and Thrap1, which are involved in cardiac hypertrophy, and connective tissue growth factor (CTGF), related to the development of fibrosis and cardiac remodeling. In 2014, Ferreira et al. published the first description of miRNA expression dysregulation in diseased myocardium of CCC patients (Ferreira et al. 2014).

The most important finding was that five muscle-specific miRNAs, miR-1, miR-133a-2, miR-133b, and the myocardial-specific miR-208a and miR-208b were downregulated in CCC myocardium as compared to control myocardium. Importantly, this study identified putative targets of the differentially expressed microRNA using a computation analysis.

They identified 2226 mRNA transcripts as putative targets of these five miRNAs tested, of which 221 had already been experimentally validated as targets. To have a preliminary assessment whether myocardial expression patterns of the 5 miRNAs were associated with concordant (i.e., inverse) expression of target miRNAs in the same tissue, they tested mRNA target matches from the gene expression microarray profiling done by Cunha-Neto et al. in 2005. Among 91 mRNAs whose expression was upregulated in CCC myocardium, 11 were targets of the concordantly downregulated miRNAs tested; they also found 3 mRNA targets that were upregulated only in DCM (out of 47) and 3 target mRNAs upregulated simultaneously both in CCC and DCM (out of 31 genes).

From the gene targets regulated by theses miRNAs there are one transcription factor, i.e., the inflammatory transcription factor and a known mediator in cardiac dysfunction, NF-κB and protein kinases, i.e., mitogen-activated protein kinases (MAPK) including p38MAPK, ERK1/2, c-Jun N-terminal kinases (JNK), phosphatidylinositide 3-kinases (PI3K), and the protein kinase B (AKT), enzymes that play important roles in signaling pathways leading to cardiac hypertrophy. Another important gene, direct target of miR-1 is cyclin D1 (CDND1). This protein, along with other D-type cyclins (D2 and D3), is a positive cell cycle regulator that plays an important role in controlling proliferation of cardiomyocytes during normal heart development. Importantly, the expression of D-type cyclins is generally low in the adult heart and is increased in the diseased heart, where their upregulation may promote cardiac hypertrophy instead of cell proliferation (Hotchkiss et al. 2012). Accordingly, a previous study has shown that CCND1 expression is upregulated during *T. cruzi* acute infection in mice and that the expression of CDND1 and other types of cyclins like A1, B1, and E1 are increased in heart lysates of mice acutely infected with *T. cruzi* compared with uninfected controls (Nagajyothi et al. 2006). The study showed that miR-1 controlled CDND1 might also be a key element in CCC.

18.9 MicroRNA Transcriptome Profiling in Heart of *T. cruzi*-Infected Mice

In 2015, our group performed microRNA transcriptome profiling in the heart of mice acutely infected with the high virulent Colombiana *T. cruzi* strain for 15, 30, and 45 days (Navarro et al. 2015). The technology used to screen 641 rodent miRNAs was a medium-throughput method, based on real-time RT-PCR that uses a set of two 384-well microfluidics cards developed by Thermo Fisher Scientific Inc., MA, USA.

The advantage of this technology is the use of stem-loop structured primers specific for binding mature miRNAs resulting in a combination of miRNA discovery and validation (Chen et al. 2005). Although this study focused on the acute phase of the experimental Chagas disease, some miRNAs (miR-133, miR-208) were found downregulated at 45 days postinfection in accordance with the previously publication reported in myocardium of CCC patients (Ferreira et al. 2014).

The use of commercial and freely available web-based computational tools (Koumakis et al. 2016) was essential in this study. The commercial software Ingenuity Pathways Analysis (IPA) (Qiagen, USA – www.ingenuity.com) which relies on three popular algorithms (TargetScan, TarBase and miRecords) was used to identify putative targets of six differentially expressed miRNAs (DEMs), selected from a list of 113 out of 641 miRNAs with significantly altered expression upon infection in at least one time point. These six miRNAs were differentially expressed in all three time-points postinfection and had the highest correlation significance with clinical parameters: such as blood parasitemia and prolongation of ventricular depolarization and repolarization time or QTc interval measured in the infected mice hearts: MiR-146b, miR-21, miR-142-3p miR-142-5p (positive correlation) and miR-145-5p and miR-149-5p (negative correlation).

Computational analysis revealed how these miRNAs potentially influence the electrocardiogram parameters: miR-149-5p has a calcium (CACNA1C) and two potassium channels as targets: KCNA1 (Kodirov et al. 2004; Newton-Cheh et al. 2009; Roder et al. 2014) and RNF207 (Roder et al. 2014); these last ones are related to repolarization of the cardiac action potential, and associated with the regulation of humans QTc interval (Gollob et al. 2006).

These miRNA profiles were also used in 2017, now to perform the first integration analysis between miRNAs and their mRNA targets in experimental Chagas heart disease using global microRNA and mRNA expression profiling from the same samples (Ferreira et al. 2017). Gene expression profiling was done using the SurePrint G3 mouse Gene Expression v1 8 × 60 K arrays, the Low Input Quick Amp Labeling One-Color Kit (Agilent Technologies, USA). By integrating these large data sets we could reveal enrichment in biological processes and pathways associated with immune response and metabolism.

Pathways, functional and upstream regulator analysis of the intersections between predicted targets of differentially expressed microRNAs and differentially expressed mRNAs revealed enrichment in biological processes and pathways such as IFNγ, TNFα, NF-kB signaling signatures, CTL-mediated apoptosis, mitochondrial dysfunction, and Nrf2-modulated antioxidative responses. We also observed enrichment in other key heart disease-related clinical outcomes, myocarditis, fibrosis, hypertrophy, and arrhythmia.

A recent study performing the same workflow integrated miRNome and transcriptome from myocardial tissue of CCC patients employing pathways and network analyses (Laugier et al. 2020). The intersections between differentially expressed microRNAs and differentially expressed target mRNAs showed that even a small number of differentially expressed microRNAs targeted a high number of differentially expressed mRNAs in multiple processes and key CCC clinical parameters like fibrosis, hypertrophy, myocarditis, and arrhythmia.

All these data collected over the years have been unveiling the players of the pathological processes in Chagas disease, like the miRNAs which revealed as key in orchestrate gene expression which bears pathogenesis, biomarkers, and therapy.

References

Abel LC, Rizzo LV, Ianni B, Albuquerque F, Bacal F, Carrara D, Bocchi EA, Teixeira HC, Mady C, Kalil J, Cunha-Neto E (2001) Chronic Chagas' disease cardiomyopathy patients display an increased IFN-gamma response to Trypanosoma cruzi infection. J Autoimmun 17(1):99–107

Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. elife 4

Bilate AM, Cunha-Neto E (2008) Chagas disease cardiomyopathy: current concepts of an old disease. Rev Inst Med Trop Sao Paulo 50(2):67–74

Bilate AM, Teixeira PC, Ribeiro SP, Brito T, Silva AM, Russo M, Kalil J, Cunha-Neto E (2008) Distinct outcomes of Trypanosoma cruzi infection in hamsters are related to myocardial parasitism, cytokine/chemokine gene expression, and protein expression profile. J Infect Dis 198(4):614–623

Bostjancic E, Zidar N, Stajner D, Glavac D (2010) MicroRNA miR-1 is up-regulated in remote myocardium in patients with myocardial infarction. Folia Biol (Praha) 56(1):27–31

Campos MA, Closel M, Valente EP, Cardoso JE, Akira S, Alvarez-Leite JI, Ropert C, Gazzinelli RT (2004) Impaired production of proinflammatory cytokines and host resistance to acute infection with Trypanosoma cruzi in mice lacking functional myeloid differentiation factor 88. J Immunol 172(3):1711–1718

Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33(20):e179

Chen JF, Mandel EM, Thomson JM, Wu Q, Callis TE, Hammond SM, Conlon FL, Wang DZ (2006) The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. Nat Genet 38(2):228–233

Coura JR, Borges-Pereira J (2010) Chagas disease: 100 years after its discovery. A systemic review. Acta Trop 115(1-2):5–13

Cunha-Neto E, Dzau VJ, Allen PD, Stamatiou D, Benvenutti L, Higuchi ML, Koyama NS, Silva JS, Kalil J, Liew CC (2005) Cardiac gene expression profiling provides evidence for cytokinopathy as a molecular mechanism in Chagas' disease cardiomyopathy. Am J Pathol 167(2):305–313

de Souza W, de Carvalho TM, Barrias ES (2010) Review on Trypanosoma cruzi: host cell interaction. Int J Cell Biol 2010

Delaporte F (1997) Romana's sign. J Hist Biol 30(3):357–366

Dias JC (1997) Cecilio Romana, Romana's sign and Chagas' disease. Rev Soc Bras Med Trop 30(5):407–413

Dias JC (2009) Elimination of Chagas disease transmission: perspectives. Mem Inst Oswaldo Cruz 104(Suppl 1):41–45

Diaz JH (2007) Chagas disease in the United States: a cause for concern in Louisiana? J La State Med Soc 159(1):21–23. 25–29

Divakaran VG (2010) MicroRNAs miR-1, -133 and -208: same faces, new roles. Cardiology 115(3):172–173

Ferreira LR, Silva AM, Michailowsky V, Reis LF, Gazzinelli RT (1999) Expression of serum amyloid A3 mRNA by inflammatory macrophages exposed to membrane glycoconjugates from Trypanosoma cruzi. J Leukoc Biol 66(4):593–600

Ferreira L, Abrantes E, Rodrigues C, Caetano B, Cerqueira G, Salim A, Reis L, Gazzinelli R (2002) Identification and characterization of a novel mouse gene encoding a Ras-associated guanine nucleotide exchange factor: expression in macrophages and myocarditis elicited by Trypanosoma cruzi parasites. J Leukoc Biol 72(6):1215–1227

Ferreira LR, Frade AF, Santos RH, Teixeira PC, Baron MA, Navarro IC, Benvenuti LA, Fiorelli AI, Bocchi EA, Stolf NA, Chevillard C, Kalil J, Cunha-Neto E (2014) MicroRNAs miR-1, miR-133a, miR-133b, miR-208a and miR-208b are dysregulated in Chronic Chagas disease Cardiomyopathy. Int J Cardiol 175 (9): 409-417.

Ferreira LRP, Ferreira FM, Laugier L, Cabantous S, Navarro IC, Cândido D d S, Rigaud VC, Real JM, Pereira GV, Pereira IR, Ruivo L, Pandey RP, Savoia M, Kalil J, Lannes-Vieira J, Nakaya H, Chevillard C, Cunha-Neto E (2017) Integration of miRNA and gene expression profiles suggest a role for miRNAs in the pathobiological processes of acute Trypanosoma cruzi infection. Sci Rep 7(1):17990

Garg N, Popov VL, Papaconstantinou J (2003) Profiling gene transcription reveals a deficiency of mitochondrial oxidative phosphorylation in Trypanosoma cruzi-infected murine hearts: implications in chagasic myocarditis development. Biochim Biophys Acta 1638(2):106–120

Gazzinelli RT, Denkers EY (2006) Protozoan encounters with Toll-like receptor signalling pathways: implications for host parasitism. Nat Rev Immunol 6(12):895–906

Goldenberg RC, Iacobas DA, Iacobas S, Rocha LL, da Silva de Azevedo Fortes F, Vairo L, Nagajyothi F, de Carvalho ACC, Tanowitz HB, Spray DC (2009) Transcriptomic alterations in trypanosoma cruzi-infected cardiac myocytes. Microbes Infect 11(14-15):1140–1149

Golgher D, Gazzinelli RT (2004) Innate and acquired immunity in the pathogenesis of Chagas disease. Autoimmunity 37(5):399–409

Gollob MH, Jones DL, Krahn AD, Danis L, Gong XQ, Shao Q, Liu X, Veinot JP, Tang AS, Stewart AF, Tesson F, Klein GJ, Yee R, Skanes AC, Guiraudon GM, Ebihara L, Bai D (2006) Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. N Engl J Med 354(25):2677–2688

Hotchkiss A, Robinson J, MacLean J, Feridooni T, Wafa K, Pasumarthi KB (2012) Role of D-type cyclins in heart development and disease. Can J Physiol Pharmacol 90(9):1197–1207

Hotez PJ, Dumonteil E, Woc-Colburn L, Serpa JA, Bezek S, Edwards MS, Hallmark CJ, Musselwhite LW, Flink BJ, Bottazzi ME (2012) Chagas disease: “the new HIV/AIDS of the Americas”. PLoS Negl Trop Dis 6(5):e1498

Junqueira C, Caetano B, Bartholomeu DC, Melo MB, Ropert C, Rodrigues MM, Gazzinelli RT (2010) The endless race between Trypanosoma cruzi and host immunity: lessons for and beyond Chagas disease. Expert Rev Mol Med 12:e29

Kodirov SA, Brunner M, Nerbonne JM, Buckett P, Mitchell GF, Koren G (2004) Attenuation of I(K,slow1) and I(K,slow2) in Kv1/Kv2DN mice prolongs APD and QT intervals but does not suppress spontaneous or inducible arrhythmias. Am J Physiol Heart Circ Physiol 286(1):H368–H374

Koumakis L, Potamias G, Tsiknakis M, Zervakis M, Moustakis V (2016) Integrating microarray data and GRNs. Methods Mol Biol 1375:137–153

Kropf SP (2011) Carlos Chagas: science, health, and national debate in Brazil. Lancet 377(9779):1740–1741

Kropf SP, Sa MR (2009) The discovery of Trypanosoma cruzi and Chagas disease (1908–1909): tropical medicine in Brazil. Hist Cienc Saude Manguinhos 16(Suppl 1):13–34

Laugier L, Ferreira LRP, Ferreira FM, Cabantous S, Frade AF, Nunes JP, Ribeiro RA, Brochet P, Teixeira PC, Santos RHB, Bocchi EA, Bacal F, Cândido DDS, Maso VE, Nakaya HI, Kalil J, Cunha-Neto E, Chevillard C (2020) miRNAs may play a major role in the control of gene expression in key pathobiological processes in Chagas disease cardiomyopathy. PLoS Negl Trop Dis 14(12):e0008889

Lescure FX, Le Loup G, Freilij H, Develoux M, Paris L, Brutus L, Pialoux G (2010) Chagas disease: changes in knowledge and management. Lancet Infect Dis 10(8):556–570

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. Cell 115(7):787–798

Machado FS, Jelicks LA, Kirchhoff LV, Shirani J, Nagajyothi F, Mukherjee S, Nelson R, Coyle CM, Spray DC, de Carvalho AC, Guan F, Prado CM, Lisanti MP, Weiss LM, Montgomery SP, Tanowitz HB (2012) Chagas heart disease: report on recent developments. Cardiol Rev 20(2):53–65

Moncayo A (2010) Carlos Chagas: biographical sketch. Acta Trop 115(1-2):1–4

Moncayo A, Ortiz Yanine MI (2006) An update on Chagas disease (human American trypanosomiasis). Ann Trop Med Parasitol 100(8):663–677

Mukherjee S, Belbin TJ, Spray DC, Iacobas DA, Weiss LM, Kitsis RN, Wittner M, Jelicks LA, Scherer PE, Ding A, Tanowitz HB (2003) Microarray analysis of changes in gene expression in a murine model of chronic chagasic cardiomyopathy. Parasitol Res 91(3):187–196

Nagajyothi F, Desruisseaux M, Bouzahzah B, Weiss LM, Andrade Ddos S, Factor SM, Scherer PE, Albanese C, Lisanti MP, Tanowitz HB (2006) Cyclin and caveolin expression in an acute model of murine Chagasic myocarditis. Cell Cycle 5(1):107–112

Navarro IC, Ferreira FM, Nakaya HI, Baron MA, Vilar-Pereira G, Pereira IR, Goncalves Silva AM, Real JM, De Brito T, Chevillard C, Lannes-Vieira J, Kalil J, Cunha-Neto E, Pinto Ferreira LR (2015) MicroRNA transcriptome profiling in heart of Trypanosoma cruzi-infected mice: parasitological and cardiological outcomes. PLoS Negl Trop Dis 9: e0003828.

Newton-Cheh C, Eijgelsheim M, Rice KM, de Bakker PI, Yin X, Estrada K, Bis JC, Marciante K, Rivadeneira F, Noseworthy PA, Sotoodehnia N, Smith NL, Rotter JI, Kors JA, Witteman JC, Hofman A, Heckbert SR, O'Donnell CJ, Uitterlinden AG, Psaty BM, Lumley T, Larson MG, Stricker BH (2009) Common variants at ten loci influence QT interval duration in the QTGEN study. Nat Genet 41(4):399–406

Nogueira LG, Santos RH, Ianni BM, Fiorelli AI, Mairena EC, Benvenuti LA, Frade A, Donadi E, Dias F, Saba B, Wang HT, Fragata A, Sampaio M, Hirata MH, Buck P, Mady C, Bocchi EA, Stolf NA, Kalil J, Cunha-Neto E (2012) Myocardial chemokine expression and intensity of myocarditis in Chagas cardiomyopathy are controlled by polymorphisms in CXCL9 and CXCL10. PLoS Negl Trop Dis 6(10):e1867

Oliveira-Carvalho V, Carvalho VO, Silva MM, Guimaraes GV, Bocchi EA (2012) MicroRNAs: a new paradigm in the treatment and diagnosis of heart failure? Arq Bras Cardiol 98(4):362–369

Pays JF (2009) Chagas Carlos Justiniano Ribeiro (1879–1934). Bull Soc Pathol Exot 102(5):276–279

Roder K, Werdich AA, Li W, Liu M, Kim TY, Organ-Darling LE, Moshal KS, Hwang JM, Lu Y, Choi BR, MacRae CA, Koren G (2014) RING finger protein RNF207, a novel regulator of cardiac excitation. J Biol Chem 289(49):33730–33740

Roveda JM (1967) Romana's sign. Cole's unilateral trypanosomiasic ophthalmia. Arch Oftalmol B Aires 42(1):1–4

Siqueira-Batista R, Gomes AP, Rocas G, Cotta RM, Rubiao EC, Pissinatti A (2011) Chagas's disease and deep ecology: the anti-vectorial fight in question. Cien Saude Colet 16(2):677–687

Soares MB, de Lima RS, Rocha LL, Vasconcelos JF, Rogatto SR, dos Santos RR, Iacobas S, Goldenberg RC, Iacobas DA, Tanowitz HB, de Carvalho AC, Spray DC (2010) Gene expression changes associated with myocarditis and fibrosis in hearts of mice with chronic chagasic cardiomyopathy. J Infect Dis 202(3):416–426

Vaena de Avalos S, Blader IJ, Fisher M, Boothroyd JC, Burleigh BA (2002) Immediate/early response to Trypanosoma cruzi infection involves minimal modulation of host cell transcription. J Biol Chem 277(1):639–644

WHO Expert Committee (2002) Control of Chagas disease. World Health Organ Tech Rep Ser 905: i–vi, 1–109, back cover

Concluding Remarks and Perspectives

I hope the second edition of *Transcriptomics in Health and Disease* has served as an updated overview of transcriptomics, from the fundamental concepts and methodology, its use in health and human disease, to the interpretation of the results. After completing the human genome, transcriptomics has entered a new realm of research, including research traditionally conducted via reductionist approaches, in particular, immunology (Chaussabel and Baldwin 2014). Because the transcriptome of a cell, tissue, or organ changes according to the strict conditions set forth at any given moment, the study of the transcriptome holds tremendous promise for health and disease research. Even disciplines that rarely adopt genetic approaches, such as physiology or pharmacology, are now examining their model systems from transcriptomics. What was once an exclusive task for geneticists and molecular biologists, i.e., sequencing the genome, has been engaged mainly by transcriptomics in the post-genome era. It has opened doors for the entire biomedical research community, including mathematicians, biostatisticians, and computer scientists. These fields have contributed to the construction of algorithms, programs for data analysis, and the improvement of bioinformatics pipelines. Without these, we would be unable to interpret the enormous quantities of data generated by bench experiments (Kharchenko 2021). And of course, clinicians themselves have seen the potential of transcriptomics in diagnosis and prognosis. Unraveling the code of life no longer involves deciphering three-letter codons (as developed by scientists in the 1960s) or sequencing all three billion bp of the human genome (mid-1980-2000), but solving the human transcriptome in response to normal physiological conditions as well as different disease states. The mouse (*Mus musculus*) is often used as a model system to answer questions of human interest, which must then be validated in humans. This is another challenge of comparative transcriptomics, which, although not explicitly discussed in this book, is currently making its mark in the literature. The core concept of the central dogma of molecular biology has not changed over the last several decades. Instead, what has happened is reinterpretation of the data, such that the "dogma" can now become: genome ➔ transcriptome ➔ proteome.

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4

Geraldo A. Passos
Ribeirão Preto,
July 2021

Chaussabel D, Baldwin N (2014) Democratizing systems immunology with modular transcriptional repertoire analysis. Nature Rev Immunol 14:271–280

Kharchenko PV (2021) The triumphs and limitations of computational methods for scRNA-seq. Nat Methods 18:723-732. Erratum in: Nat Methods. 2021 June 30

Nguyen LV, Caldas C (2021) Functional genomics approaches to improve pre-clinical drug screening and biomarker discovery. EMBO Mol Med 13:e13189

Index

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4