Human Genome Structure, Function and Clinical Considerations

Luciana Amaral Haddad
Editor

Springer

Human Genome Structure, Function and Clinical Considerations

Luciana Amaral Haddad
Editor

Human Genome Structure, Function and Clinical Considerations

Springer

Editor
Luciana Amaral Haddad
Department of Genetics and
Evolutionary Biology
Instituto de Biociências
Universidade de São Paulo
São Paulo
Brazil

ISBN 978-3-030-73150-2 ISBN 978-3-030-73151-9 (eBook)
https://doi.org/10.1007/978-3-030-73151-9

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

Preface

Since the publication of the first draft of the human genome sequence twenty years ago, the genomics field has experienced rapid development coupled with technological breakthroughs mostly based on novel DNA sequencing platforms and advances in bioinformatics. This book has been organized to present the human genome components under an updated functional view by specialists in the field while highlighting certain clinical correlations to genomic variation and the major historical milestones that contributed to the basis and progress of genomic medicine.

It has been an honor to have Prof. Sérgio D. Pena as a distinguished guest author giving an overview of the human genome in the first chapter, based on his vast experience as medical geneticist and professor of molecular biology. The reader may refer to this chapter for the ensemble of genomic elements, which are individually covered in consecutive sections of the book. I am grateful to the colleagues who contributed with chapters and are accomplished researchers in various fields of human genetics.

Chapters 2 and 3 provide a thorough view of the structure of human chromosomes, the molecular basis of inheritance, and the methods currently employed to assess chromosomal structural and numerical alterations and sequence variation in the human genome DNA. Chapters 4 and 5 present the diversity of protein-coding and noncoding genes, their functional relevance and the elements regulating their expression. The distinct classes of human genomic DNA repeats are discussed in Chaps. 6–8, whereas segmental DNA copy number variations are introduced in Chap. 9. The unique characteristics of the human mitochondrial DNA are presented in Chap. 10, while Chap. 11 addresses genome-wide variations in human populations, as well as the major evolutionary factors that justify their frequencies. Importantly, information of clinical significance based on scientific evidence related to a discussed genomic concept illustrates the respective chapter.

I would like to acknowledge those that made possible the organization and publication of this book. I thank the Springer Nature publisher through the executive editor, Grant Weston, the book project coordinator Anand Shanmugan, and the production team. I also recognize the support and incentive by Prof. Dhavendra Kumar (Queen Mary University of London) on the first steps of this project. Working as

researcher in a higher education environment is a two-way road that allows the research work feed the teaching activities and backwards. Thus, I particularly thank the University of São Paulo for the academic setting and São Paulo Research Foundation (FAPESP) for funding research activities in my laboratory (currently through processes 2013/08028-1 and 2019/10868-4). The thoughtful classroom questioning by undergraduate and graduate students is a motivation for our continuous learning. I am indebted to the many students of the molecular biology courses I have taught and my colleagues with whom I have the pleasure to work with.

The latest developments in human genome analysis have extended the reach of genetics and genomics to different professionals in health care. Genomics hence permeates every specialty in medicine. It is our expectation that the *Human Genome Structure, Function, and Clinical Considerations* book, by providing direct access to the genomics terminology and state-of-the-art human genome information, serves as a resourceful material for clarification, reviewing, and consulting to established health care professionals as well as students initiating in the biomedical field.

Sincerely

São Paulo, Brazil Luciana Amaral Haddad

Contents

Chapter 1
An Overview of the Human Genome

Sérgio D. J. Pena

1.1 General Introduction

The term “Genetics” was coined in 1906, by William Bateson, the great defender of Mendel’s ideas in England and the first to propose autosomal recessive inheritance for a human disease, alkaptonuria. Paradoxically, the name “gene” was only used 3 years later by Wilhelm Johannsen, a Danish botanist, to indicate an abstract unit of inheritance. Johannsen had no idea what a gene could be. It was just a theoretical concept that allowed him to understand the phenomena of heredity of discontinuous characters and could even explain the appearance of some human diseases. This situation lasted for many years. With the emergence of the chromosomal theory of heredity, the gene came to be considered a dimensionless point within a carrier chromosome. According to this view, the genes were lined up like beads from a rosary. The interesting thing is that even though scientists had not yet found a chemical identity for genes, genetics flourished and became a respected branch of science, with thousands of articles and hundreds of books published.

It was only in 1953 that, with the triumphant announcement of James Watson and Francis Crick, that the gene gained a bodily structure in the double helix of DNA. Then, a new experimental science was born, molecular biology, which would largely replace the predominantly statistical techniques of classical genetics.

Since 1953 we came to understand genes as DNA segments responsible for encoding a genetic trait, a polypeptide, in general by means of a functional RNA. The genes are found aligned in segments within the chromosomes, which, in turn, are a framework made up of long DNA molecules in association with proteins.

S. D. J. Pena (✉)
Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais,
Belo Horizonte, MG, Brazil

GENE – Núcleo de Genética Médica, Belo Horizonte, MG, Brazil
e-mail: spena@dcc.ufmg.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_1

The next step was to map which genes were responsible for different characteristics and in what order they were on the different chromosomes within the cell nucleus. In other words, we wanted to establish a map of all the DNA in all chromosomes—this set has been named “genome”. This mapping was done by the Human Genome Project, which started in 1989 and ended solemnly on April 25, 2003, commemorating the 50th anniversary of the discovery of Watson and Crick. In that date, we officially entered the Genomic Era.

The term genome, however, was not born in 1989. It had been proposed in 1920 by the plant geneticist Hans Winkler to describe “the haploid chromosome set”, which, together with the pertinent protoplasm, specifies the material foundations of the species” (reviewed in [1]). The term really only took off recently, with our understanding of the architecture of genomes, of the evolution of genomes and of the genetic expression of the genome. Even so, as pointed out by Goldman and Landweber [1], today the standard definition of the genome remains very similar to its 1920 meaning. At the Genetics Home Reference website of the National Institutes of Health (NIH) the genome is defined as: “A genome is an organism’s complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism. In humans, a copy of the entire genome—more than 3 billion DNA base pairs—is contained in all cells that have a nucleus” [2].

Under this definition, a genome may seem static and permanent. Much on the contrary, genomes, especially those of somatic cells, are always changing, in constant turnover. In fact, such changes underlie the generation of antibody diversity and the so-called “genomic disorders” of somatic human tissues: cancer, autoimmune diseases and, perhaps, aging itself. A relevant example of the physical impermanence of the human genome occurs in the retroviral infection cycle. Upon infection, retroviruses convert their single-stranded RNA genomes into double-stranded DNA. These intermediate DNA copies of the genome are integrated into the host cell and, thus, no longer constitute a separate physical entity from the host’s genome. As an integrated DNA sequence, transcription into mRNA can both express retroviral genes and also reconstitute the original single-stranded (ss) RNA genome [1]. We will see more about retroviruses, endogenous and exogenous, in later sections of this text.

The science of studying genomes is called “genomics”. How do human genetics and human genomics differ? Human genetics involves the study of limited groups of genes in a specific individual or in defined populations. It is possible to study genetics without knowing the biochemical characteristics of the genes, as was done from 1900 to 1953, a period in which a gene was simply identified through the manifestation of a disease or some characteristic that can be observed segregating in a pedigree. On the other hand, human genomics is the study of the totality of the DNA of all the chromosomes in *Homo sapiens*. Using high-performance computing and mathematical techniques (bioinformatics), researchers in genomics analyze massive amounts of sequence data of DNA to find variations that affect health and are associated with diseases.

This chapter was initially designed to present a brief introduction to the human genome and human genomics—a "CliffsNote view". One could write a lot about the human genome and, indeed, the subject has been covered in whole books. The textbook "Genomes 4", for instance, has 544 pages [3] and the textbook "Genetics and Genomics in Medicine" has 524 pages [4]. The reader is directed to them for more complete information.

To avoid a lengthy chapter that would certainly meet the disapproval, if not the scissors, of the editor of this book, I have chosen only three main topics in which to concentrate my writing: (A) a bird's-eye view of selected aspects of the structure and evolution of the human genome; (B) the consequence of the abundance of retroposons, retrotransposons and endogenous retroviruses in the human genome, and (C) genomic sequencing as a tool for the accurate prediction of disease risk in Precision Medicine. The latter is a subject of great current interest, since the utilization of genomic sequencing for the accurate prediction of disease risk is a necessity for the development of Precision Medicine.

Within each of these topics I have not written a coherent textbook-like presentation. If the reader is interested in facts and hard data, the Wikipedia article on Human Genome should be consulted. Rather than trying to be comprehensive, I chose instead to discuss matters of my interest, occasionally employing a lighter vein.

1.2 A Bird's-Eye View of Selected Aspects of the Structure and Evolution of the Human Genome

1.2.1 Introduction

The human genome, with 3.2 billion base pairs is, by definition, haploid (n), constituting the genetic material present in a single human gamete. It is composed of 22 autosomal chromosomes, numbered 1–22 (autosomes), and a sex chromosome, called X or Y. The haploid chromosomal set of an ovum and a spermatozoon will join to form the zygote (precursor of somatic cells), which is diploid and has 46 chromosomes, to wit, 44 autosomes and two sex chromosomes (XX in females, XY in males). Every gene in autosomes is present in two copies in the zygote. Thus, the zygote and descendant somatic cells contain not one, but two genomes (one maternal and one paternal). After the first mitotic division of the zygote these two genomes get mixed together in the cell nucleus.

The human genome is made up of a panoply of different DNA components, as shown in Fig. 1.1, which was produced by the NHS National Genetics and Genomics Education Centre in 2014 [5]. It is always surprising to observe that only circa 2% of the human genome is composed of protein-coding genes.

The total number of structural genes (protein-coding) distributed on the 23 chromosomes of the human genome is estimated to be 20,412, slightly less than the

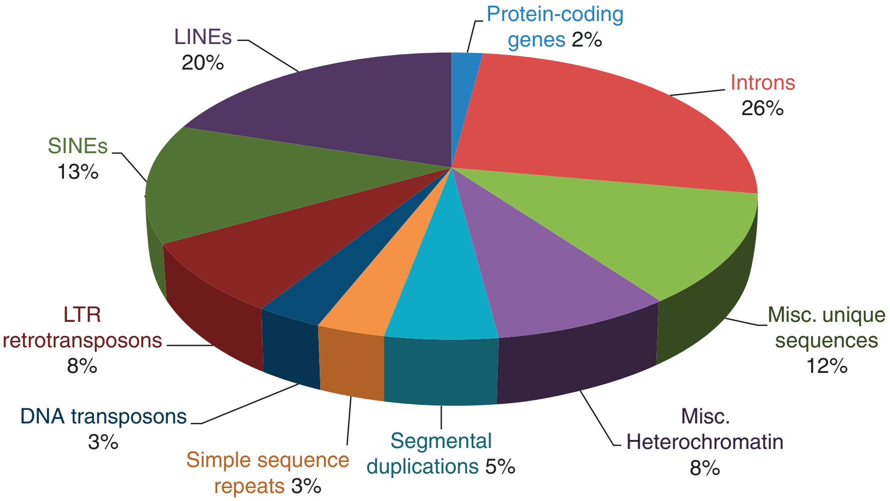


Fig. 1.1 Composition of the human genome. Redrawn from a graph that was produced in 2014 by the NHS National Genetics and Genomics Education Centre [5]

20,470 genes of the nematode *Caenorhabditis elegans*! For humans, accustomed to feel that they are at the top of the biological scale, this finding is unflattering. However, a mechanism called "alternative editing" makes it possible to multiply these genes into hundreds of thousands of different transcripts, some of them tissue-specific.

1.2.2 The Basic Morphological Division of the Human Genome: Chromosomes

The number of genes on each human chromosome varies widely, ranging from 2058 genes on chromosome 1 to only 71 genes on Y chromosome. The three autosomes with the fewest genes are chromosome 13 (327 genes), chromosome 18 (270 genes) and chromosome 21 (234 genes). It is thus no accident that the only autosomal human trisomies compatible with the survival of the fetus till birth are trisomies 13, 18 and 21! The density of genes on chromosomes also varies widely. For instance, chromosome 19 is smaller than chromosome 13, but contains almost four times more genes than the latter (chromosome 19 is the second in decreasing order of gene content, just behind the chromosome 1).

There is apparently no specific reason why humans have 46 chromosomes in somatic cells. Our closest primate, the chimpanzee (*Pan troglodytes*) has 48 chromosomes. In the evolution of primates, two acrocentric chromosomes from the chimpanzee underwent centric fusion to form human chromosome 2, hence the reduction of chromosome number to 46. In contrast, the mouse (*Mus musculus*) has

56 chromosomes. The *Lysandra atlantica* butterfly has 446 chromosomes in diploid cells, while *Lysandra golga* has 268 and *Lysandra nivescens* has 82! In fact, there seems to be no correlation between the number of chromosomes or the size of the total genome or the biological complexity of the species. Both seem to vary at random. Thus, everything suggests that the chromosomes may be only physical frameworks that allow the realization of mitosis and meiosis in sexual species [6].

In chromosomes, DNA contains genes that are expressed according to the needs of the cell, but it also contains specialized sequences that are necessary for intrinsic functions of the chromosome itself. On one hand, chromosomes need to be properly aligned during cell division. This requires a centromere, a region where a pair of protein complexes, called kinetochores, binds just before the start of cell division. Microtubules are responsible initially for positioning the chromosomes correctly in the metaphase and then for pulling the individualized chromosomes to opposite poles of the mitotic spindle. The DNA sequences in the centromeres are very different in different organisms. In mammalian chromosomes, centromeric DNA is a heterochromatic region, with no informational content, dominated by repetitive DNA sequences that often extend monotonously by mega DNA bases.

At the ends of chromosomes, there are specialized structures called telomeres, which are necessary for maintaining chromosomal integrity. If a telomere is lost after a break in a chromosome, the resulting chromosomal end is unstable and tends to merge with the broken ends of the other chromosomes, or even be degraded. In vertebrate telomeres the DNA consists of multiple copies in tandem of the oligonucleotide TTAGGG, sequence at which certain telomeric proteins bind. The repetitive units of the telomeres decrease in number with every division of the DNA. As the enzyme needed to regenerate telomeres (telomerase) is not available in normal somatic cells, telomeres are a kind of biological clock that records our age.

The chromosomes seem to behave functionally as "packages" of genes. In general, the functioning of individual genes is not affected by their chromosomal position. For instance, there are individuals with balanced chromosomal translocations, in which chromosomes have exchanged segments without loss or net gain of genetic material—such individuals do not present any clinical manifestation of translocation, except perhaps for reproductive difficulties, as some translocations may interfere with the production of gametes in meiosis, especially in the male.

1.2.3 Coding DNA and Non-coding DNA

As we have already seen, genes are the segments of DNA that carry genetic information to produce proteins or functional RNA molecules. The vast majority of genes are in the chromosomes of the nucleus; a few are also found in mitochondrial DNA. Remarkable similarities of known human and chimpanzee protein sequences initially led to the suggestion that significant differences might be primarily in gene and protein expression, rather than protein structure. Further analysis of alignable non-coding sequences affirmed this ~1% difference. However, the subsequent identification of non-alignable sequences that were due to segmental deletions and duplications has shown that the overall difference between the two genomes is actually ~4% [7].

Less than 2% of the human genome corresponds to protein-coding genes (Fig. 1.1). The functional role of the remaining 98%, apart from repetitive sequences (constitutive heterochromatin) that appear to have a structural role in the chromosome, is a matter of controversy. Evolutionary evidence suggests that this noncoding DNA has no function—hence the common name of "junk DNA". The most convincing such evidence is the so-called Paradox of Value C (C-value is the amount in picograms of the DNA contained within a haploid nucleus, thus being basically synonymous to genome size). The paradox is that across evolutionary lines of eukaryotes, the genome size shows no correlation with the number of genes and the biological complexity of the species [6]. The part of the genome that is responsible for the paradox is exactly the non-coding part, which varies 300,000-fold among different species, even sometimes closely relates ones. There is an amoeba, which has a genome 214 times greater than the human genome! As a percentage of the genome size the non-coding portion varies from less than 30% to 99.998% of the total genome! [6].

Graur et al. [8] mention the "Onion test" that apparently was originally stated by the Canadian evolutionary biologist T. Ryan Gregory: The onion test is a simple reality check for anyone who thinks they can assign a function to every nucleotide in the human genome. Whatever your proposed functions are, ask yourself this question: Why does an onion need a genome that is about five times larger than ours?"

Since no obvious function could be assigned to the non-coding genome and it is largely irrelevant to fitness, it was called "junk DNA". Here we have to delve a bit into semantics: "junk" is not "garbage"! According to the Merriam-Webster dictionary (Merriam-Webster.com) "junk" is "old iron, glass, paper, or other waste that may be used again in some form", while "garbage" is "food waste; discarded or useless material". In other words, junk is something that you keep and garbage is something that you discard.

The "junk DNA" hypothesis seemed to become the canonical view in genome biology. However, in 2012 there was the publication of the ENCODE Project, communicating very different findings [9]. I quote from their summary: "The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research." What they called "biochemical functions" included primarily the annotation of 8801 automatically derived small RNAs and 9640 manually curated long non-coding RNA (lncRNA) loci. The ENCODE project also annotated 11,224 pseudogenes, of which 863 were transcribed and associated with active chromatin.

The ENCODE project's claim that 80% of the human genome has biochemical function raised a semantic discussion of the meaning of "functional". Clearly the ENCODE project used a very liberal definition of "functional": anything that is transcribed must be functional. This is clearly not always the case. For instance, although some pseudogenes may be transcribed, they are nevertheless non-functional. Moreover, we still do not have a clear notion of the potential function of these long non-coding RNAs that ENCODE disclosed [10].

Many evolutionary biologists have stuck to their guns in defense of the traditional genetic and evolutionary view that non-coding DNA is "junk DNA". I will quote from Graur et al. [8]: "The recent slew of ENCyclopedia Of DNA Elements (ENCODE) Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. This claim flies in the face of current estimates according to which the fraction of the genome that is evolutionarily conserved through purifying selection is less than 10%. Thus, according to the ENCODE Consortium, a biological function can be maintained indefinitely without selection, which implies that at least 80–10 = 70% of the genome is perfectly invulnerable to deleterious mutations, either because no mutation can ever occur in these "functional" regions or because no mutation in these regions can ever be deleterious. This absurd conclusion was reached through various means, chiefly by employing the seldom used "causal role" definition of biological function and then applying it inconsistently to different biochemical properties, by committing a logical fallacy known as "affirming the consequent," by failing to appreciate the crucial difference between "junk DNA" and "garbage DNA", by using analytical methods that yield biased errors and inflate estimates of functionality, by favoring statistical sensitivity over specificity, and by emphasizing statistical significance rather than the magnitude of the effect."

1.2.4 Natural Selection and Genome Evolution

As pointed out by Koonin [11], Charles Darwin believed that all the characteristics of organisms were improved almost to perfection by natural selection. The empirical basis underlying Darwin's conclusions consisted of numerous observations made by him and others about the exquisite adaptations of animals and plants to their natural habitats and the impressive results of artificial selection. Now, more than two centuries after the birth of Darwin, we can compare hundreds of genome sequences from several species and draw new conclusions about evolutionary events. These comparisons suggest that the dominant mode of evolution of the genome is different from that of phenotypic evolution. The vertebrate genomes have turned out to be true junkyards of selfish genetic elements, where only a small fraction of the genetic material is dedicated to encoding biologically relevant information.

In 2009, Koonin published a fascinating review entitled "Darwinian evolution in the light of genomics" [11]. In it, he lays out six fundamental principles of Darwinism, grouping together the propositions made by Darwin and those of the Modern Synthesis). For each of the six topics, Koonin [11] provided the corresponding Darwinian proposition and the then current status of the proposition under the light of genomics. We believe that this panorama has remained basically unchanged up to now:

(a) **Darwinian statement: the material for evolution is provided, primarily, by random, heritable variation.**

True. The repertoire of relevant random changes greatly expanded to include duplication of genes, genome regions, and entire genomes; loss of genes and, generally, genetic material; horizontal gene transfer (HGT) including massive gene flux in cases of endosymbiosis; invasion of mobile selfish elements and recruitment of sequences from them; and more.

(b) **Darwinian statement: the fixation of (rare) beneficial changes by natural selection is the main driving force of evolution that, generally, produces increasingly complex adaptive features of organisms; hence progress as a general trend in evolution.**

False. Natural (positive) selection is an important factor of evolution but is only one of several fundamental forces and is not quantitatively dominant; neutral processes combined with purifying selection dominate evolution. Genomic complexity, probably evolved as a 'genomic syndrome' caused by weak purifying selection in small population and not as an adaptation. There is no consistent trend towards increasing complexity in evolution, and the notion of evolutionary progress is unwarranted.

(c) **Darwinian statement: the variations fixed by natural selection are 'infinitesimally small'. Evolution adheres to gradualism.**

False. Even single gene duplications and HGT of single genes are by no means 'infinitesimally small' let alone deletion or acquisition of larger regions, genome rearrangements, whole-genome duplication, and most dramatically, endosymbiosis. Gradualism is not the principal regime of evolution.

(d) **Darwinian statement: uniformitarianism: evolutionary processes remained, largely, the same throughout the evolution of life.**

Largely true. However, the earliest stages of evolution (anteceding the last eukaryotic common ancestor—LUCA), probably, involved distinct processes not involved in subsequent, 'normal' evolution. Major transition in evolution like the origin of eukaryotes could be brought about by (effectively) unique events such as endosymbiosis.

(e) **Darwinian statement: the entire evolution of life can be depicted as a single 'big tree'.**

False. The discovery of the fundamental contributions of HGT and mobile genetic elements to genome evolution invalidate the concept of the single tree of life (TOL) in its original sense. However, trees remain essential templates to represent evolution of individual genes and many phases of evolution in groups of relatively close organisms. The possibility of salvaging the TOL as a central trend of evolution remains.

(f) **Darwinian statement: All extant cellular life forms descend from very few, and probably, one ancestral form (LUCA).**

True. Comparative genomics leaves no doubt of the common ancestry of cellular life. However, it also yields indications that LUCA(S) might have been very different from modern cells.

1.2.5 A Borgesian View of the Human Genome

The metaphorical vision of the human genome as a library has become almost trite and commonplace. Since the early days of molecular biology, linguistic, grammatical or bibliographical images have been employed extensively. For instance, we say that the information in coding DNA (the genes), which is written in an alphabet of four letters (the bases), is transcribed into messenger RNA, and eventually translated into a protein language, which uses an alphabet of 20 letters (the amino acids) according with the rules of the genetic code. Interestingly, with the spread of the DNA meme, the converse also became true, i.e. libraries can be imagined as DNA. Take, for instance, this statement from Susan Orlean: “Books are sort of a cultural DNA, the code for who, as a society, we are and what we know” [12].

As we have seen already, before the Human Genome Project, the model that we had of the human genome was of a well-organized structure, more or less static, in which individual genes had a precise place preordained by their function. Thus, it made a lot of sense to think in terms of a library, in which the genes were the texts and the chromosomes were the shelves or sections, all having evolved under the aegis of natural selection. However, the final picture that emerged from the Human Genome Project (HGP) was completely different from this!

Our genome is more like a deposit room or attic than a library: unkempt, with no evidence of organization, full of refuse, debris and scrap (non-coding DNA)—nothing is discarded, even if it is useless at the moment. However, at any moment a new use may be found for some of the stored junk. Besides, the human genome is very dynamic, its pieces being shuffled and changed frequently without any reason or rhyme. In fact, the coding genes are scarce (less than 2% of the total!) and are scattered carelessly among enormous amounts of often highly repetitive DNA that lacks apparent function or sense, the so-called ‘junk-DNA’.

If we look at the human genome from an evolutionary point of view, comparing it with other genomes, the situation gets even more complicated. Total genome size does not mean much. The lily, the newt and many, many others have genomes much larger than ours. There is even a prosaic amoeba (*Amoeba dubia*) that has a genome with 690 billion base pairs, more than 200 times the size of the human. These size differences do not reflect a variation in the number of genes, but in the amount of non-coding DNA (junk-DNA). Likewise, as we have already seen, the chromosome number does not have much significance. Finally, the content itself lacks meaning. Three non-coding sequences, the retroposon Alu, the retrotranposons L1, and endogenous retroviruses (LTR-containing retrotransposons represent 41% of the total (Fig. 1.1)! In the kangaroo rat (*Dipodomys ordii*), more than 50% of the genome consists of only three simple repetitive sequences, one of which—AAG—is repeated more than one billion times. There is need for a lot of imagination to try to visualize design or necessity in this genomic mess. If the human genome is a library what kind of library is it?

The Argentinian *Jorge* Luis Borges (1899–1986) was a marvelous and unique writer. His stories, fantastic and enigmatic, are very short, but pack more content than whole books and induce us to spend hours in philosophical speculation. Of particular interest for us here is the short story entitled "The Library of Babel", published in the book Ficciones in 1944 [13]. This library became an integral part of the novel "The Name of the Rose", written by Umberto Eco (1932–2016), which features a labyrinthine library, presided over by a blind monk called *Jorge* of Burgos. "The Library of Babel" reminded me a lot of the human genome. I propose then a little game. I have selected some passages of Borges, that bring to mind numerous structural similarities with the human genome. Here they are:

- Also, through here passes a spiral stairway, which sinks abysmally and soars upwards to remote distances. In the hallway there is a mirror which faithfully duplicates all appearances.
- There are also letters on the spine of each book; these letters do not indicate or prefigure what the pages will say.
- This much is already known: for every sensible line of straightforward statement, there are leagues of senseless cacophonies, verbal jumbles and incoherencies.
- Four hundred and ten pages of inalterable MCV's cannot correspond to any language, no matter how dialectical or rudimentary it may be.
- Every copy is unique, irreplaceable, but (since the Library is total) there are always several hundred thousand imperfect facsimiles: works which differ only in a letter or a comma.
- The impious maintain that nonsense is normal in the Library and that the reasonable (and even humble and pure coherence) is an almost miraculous exception.
- The Library is unlimited and cyclical. If an eternal traveler were to cross it in any direction, after centuries, he would see that the same volumes were repeated in the same disorder (which, thus repeated, would be an order: the order). My solitude is gladdened by this elegant hope.

To finish, I wish to borrow a small passage from another marvelous short story by Borges, also from the book Ficciones, called "The Garden of Forking Paths": "… no one realized that the book and the labyrinth were one and the same."

1.3 Consequence of the Abundance of Retroposons, Retrotransposons and Endogenous Retrovirus in the Human Genome

Transposable elements can be separated into two major classes: DNA transposons and retroposition elements [14]. DNA transposons, which constitute approximately 3% of the human genome (Figs. 1.1 and 1.2), can excise themselves from the genome, move as DNA and insert themselves into new genomic sites. Although DNA transposons are currently not mobile in the human genome, they were apparently active during early primate evolution [14].

Retroposition elements, i.e. retroposons, retrotransposons and endogenous retroviruses, duplicate through RNA intermediates that are reverse transcribed and inserted at new genomic locations. Together, they constitute more than 40% of the human genome (Figs. 1.1 and 1.2).

The retroposons do not contain the gene for reverse transcriptase and thus are dependent on exogenous sources of the enzyme (mostly from Long Interspersed Nuclear Elements—LINEs) for retroposition. They share similarity with genes transcribed by RNA polymerase III, the enzyme that transcribes genes into ribosomal RNA, tRNA and other small RNA molecules. An especially abundant group of retroposons in humans is the Alu family of SINEs (Short Interspersed Nuclear Elements), that basically represents a processed pseudogene of the Signal Recognition Particle (7SL) RNA. The Alu family of retroposons (thus called because they contain a site for digestion by the restriction enzyme AluI) makes up 13% of the human genome. Virtually all other mammalian SINEs differ from the human, being derived from tRNA genes.

In contrast, retrotransposons do code for reverse transcriptase and hence are capable of autonomous retrotranscription. They also contain a promoter for RNA polymerase II, which allows it to insert itself into random positions. In humans, LINEs, which altogether make up 20% of the human genome, are the main class of

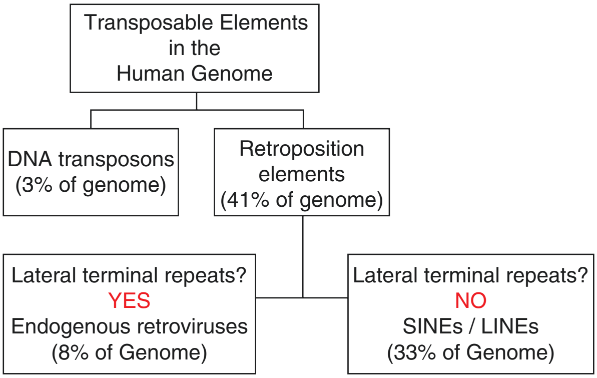


Fig. 1.2 Classes of transposable elements in the human genome

retrotransposons. Although the vast majority of human LINE-1 sequences are inactive molecular fossils, an estimated 80–100 copies per individual still retain the ability to mobilize and expand in numbers within the human genome, by cycles of transcription, retrotranscription and retroposition. Some of these active LINEs constitute insertional polymorphisms in the human species [14]. LINEs and SINEs continue growing in numbers in all mammalian genomes, and thus are "genomic parasites", the ultimate "selfish genes". For further information see Richardson et al. [15].

Finally, we have the class of retrotransposons that contain lateral terminal repeats (LTRs), better known as endogenous retroviruses, which are evolutionarily related to the exogenous retrovirus group of RNA virus and will be the focus of this section (Fig. 1.2). They constitute around 8% of the human genome! This is ironic, considering that at the very moment that I am writing this chapter humanity is being held ransom by the RNA virus SARS-CoV-2 that causes the serious disease COVID-19 [16]. Thus, if not only for its timeliness, I think that today any discussion of the structure and function of the human genome should include a discussion of these endogenous retroviruses. In special I want to evaluate the evidence for a conceivable anti-viral effect of these mostly defective and dormant endogenous retroviruses, which eons ago were exogenous, infected germ cells, endogenized and multiplied to become 8% of the human genome.

A note on nomenclature before we dive in: an endogenous retrovirus is generally called ERV or EVE (endogenous viral element). Although not one of the thousands of retrovirus-related sequences found in the human genome contains a complete set of intact retroviral genes or can express infectious virus, these sequences are nonetheless referred to as Human Endogenous Retroviruses (HERVs [17, 18]).

Viruses and/or virus-like selfish elements are associated with all cellular life forms and are the most abundant biological entities on Earth, with the number of virus particles in many environments exceeding the number of cells by one to two orders of magnitude [19]. Unlike cellular organisms with their uniform replication-expression scheme, viruses possess either RNA or DNA genomes and exploit all conceivable replication-expression strategies. We are here concerned with retroviruses, which form a family of enveloped RNA viruses entirely limited to vertebrate hosts. Exogenous RNA viruses, such as SARS-CoV-2 are transmitted horizontally among hosts. The number of described exogenous retroviruses is small, and is distributed across 7 genera, containing just 53 described species, although metagenomic studies suggest that our understanding of the true biodiversity and evolution of vertebrate RNA viruses may be fragmentary and biased [20]. On the other hand, ERVs, which are inherited vertically in the genomes of their hosts, represent a great wealth of retroviral sequence diversity that has accumulated over millions of years of vertebrate–retrovirus interactions. ERVs accumulate mutations at the background rate of sequence mutation in their host's genome, gradually degrading until their sequences are no longer recognizable [21].

HERVs share with exogenous retroviruses the typical proviral structure, being normally composed of two long terminal repeats (LTRs) that flank the internal portion of the viral genes gag, pro-pol and env. The LTRs are formed during the reverse transcription and have a regulatory significance for the expression of the viral genes' expression, including promoters, enhancers and polyadenylation signals. The retroviral genes encode the structural components, i.e. matrix, capsid and nucleocapsid (gag) and the envelope surface and transmembrane subunits (env), as well as the enzymes involved in the viral life cycle, namely protease (pro), reverse transcriptase and integrase [22]. Some HERVs still have open reading frames with possibility of protein expression. There have been 3173 HERV sequences identified from the human genome, and 39 canonical types of HERVs can be categorized as belonging to classes I, II and III on the basis of sequence similarity to different genera of exogenous retroviruses, i.e. Gammaretrovirus/Epsilonretrovirus, Betaretrovirus and Spumaretrovirus, respectively [23].

Between the 5′ LTR and gag, a primer-binding site (PBS) is located that has traditionally been used for a systematic nomenclature of HERV. HERV group names are generally identified using a letter that characterizes the human tRNA type that binds to the viral PBS sequences during the retrotranscription process of the viral genome, e.g., HERV-K for lysine, HERV-W for tryptophan, etc. [24]. Only the HERV-K (HML-2) family shows evidence of recent activity within the human genome. Although no single HERV locus has been found that can produce infectious virions, the reconstitution of an ancestral HERV-K (HML-2) genome resulted in the production of functional infectious viral particles [25].

Because of the constant genome turnover mechanisms, occasionally recombination between the 5′ and 3′ of an endogenous retrovirus results in complete loss of all viral genes and formation of a "solo LTR" at the same location that the endogenous retrovirus used to be located. Formation of solo LTRs is common, and solo LTRs in genomes often outnumber other ERV sequences by orders of magnitude [17, 18].

For most retroviruses that have been studied in detail, the primary targets of infection are cells of somatic tissues. However, since the genomes of almost all vertebrates (including humans) carry hundreds of thousands of integrated retroviral sequences, retroviral infections of organisms must have occasionally resulted in the infection of germline cells. In most cases, these are the remnants of proviruses left there by ancient exogenous retroviruses, by now possibly extinct [17]. However, the process of endogenization is not only confined to the ancient past. Recent endogenization has been documented in some species, including mice and koalas. The potential for endogenization exists anywhere that a retrovirus is spreading within a population of host organisms, although the probability of an endogenous provirus forming may be strongly influenced by the biology of the particular virus [17].

Hayward [21] asks a very relevant question: which came first, the exogenous retroviruses or the retroposition elements? Basically, the only difference between them is that the viruses have an env gene that codes for a lipoproteic envelope. LTR transposons may have evolved from a viral ancestor by losing the env gene (with subsequent gains of env-like genes in some cases). Alternatively, exogenous retroviruses might have evolved from an LTR transposon ancestor by gaining an envelope gene.

Gould and Vrba [26] coined the term exaptation to be used in reference to genetic elements that evolved under one set of selective pressure, but have later been coopted by natural selection to fulfill a different function. An interesting and well-documented case of HERV exaptation are the syncytins, proteins involved in placental implantation that have evolved from envelope proteins encoded by endogenous retroviruses on multiple occasions during mammalian evolution. In the case of the human genome, two env loci, namely *ERVW-1* (MIM*604659) and *ERVFRD-1* (MIM*610524), encode for the coopted Env proteins syncytin-1 and syncytin-2, respectively. While syncytin-1 has a pivotal role in placental syncytiotrophoblast development and homeostasis, syncytin-2 is thought to perhaps be involved in the maternal immune tolerance to the fetal allograft [22].

Also, there are now several examples of ERV-encoded proteins that have been coopted as defenses against infection by exogenous retroviruses. Relevant to this, HERVs can be potentially considered as restriction factors able to exert protective effects against exogenous retroviruses [22]. Three possible mechanisms have been suggested by which HERV could promote resistance to exogenous retrovirus infections: (1) Occurrence of complementary interactions between HERV mRNAs and homologous RNAs originated by exogenous retrovirus, with formation of dsRNA molecules that, in turn, can stimulate the Toll-Like Receptor 3 (TLR3) and thus an innate immune response [25]. (2) Aggregation of HERV and retroviral proteins, as observed in cells co-infected by both HIV-1 and HERV-K, in which gag proteins of both viruses colocalized at the plasma membrane and co-assembly into the same HIV- 1 virions, thus inhibiting release of new HIV-1 infectious particles [27]. (3) Superinfection interference, as that exerted by HERV pseudo-viral particles or HERV-derived proteins that block retrovirus entry through cellular-receptor interference. This was the case of the truncated HERV-F env protein encoded by the suppressyn gene, that by binding the cell receptor ASCT2 might prevent the entry of several type D-retroviruses [24].

However, in spite of the above *in vitro* observations, there is not yet direct evidence that the env genes found in the human genome confer resistance to retroviruses *in vivo* [17]. On the other hand, it is noteworthy that none of the three clades of HERVs (gammaretroviruses, beta-like and spuma-like retroviruses) have exogenous counterparts that are autonomously infectious for humans (spuma-like viruses only infect humans by zoonosis). This is of special significance because several oncoviruses that belong to the gammaretroviruses group (murine leukemia virus, Abelson murine leukemia virus, Friend virus, feline leukemia virus and xenotropic murine leukemia virus-related virus) are not able to infect humans. The opposite seems also to be true: exogenous retroviruses that are able to infect human cells [lentiviruses (HIV) and deltaretroviruses (HTLV-1, HTLV-2)] do not have endogenous representatives in humans. These facts are compatible with the hypothesis that HERVs have being coopted as an antiviral defense against related exogenous retroviruses [28].

1.4 Genomic Sequencing as a Clinical Tool for the Accurate Prediction of Disease Risk in Precision Medicine

1.4.1 What Is Precision Medicine?

In his famous 2015 State of the Union address, President Barack Obama [29] announced that he was launching the Precision Medicine Initiative, which was characterized as “a bold new research effort to revolutionize how we improve health and treat disease”. He further described it with the following statement: “Until now, most medical treatments have been designed for the average patient. As a result of this one-size-fits-all approach, treatments can be very successful for some patients but not for others. Precision Medicine, on the other hand, is an innovative approach that takes into account individual differences in people’s genes, environments, and lifestyles. It gives medical professionals the resources they need to target the specific treatments of the illnesses we encounter, further develops our scientific and medical research, and keeps our families healthier”.

Other names for “Precision Medicine” are “Personalized Medicine” and “Genomic Medicine”. The physician and molecular biologist Leroy Hood (who received the “National Medal of Science” from President Obama), called it “P4 Medicine”, since it is at the same time *P*redictive, *P*ersonalized, *P*reventive, and *P*articipatory. As idealized, P4 Medicine focuses on individuals, not populations. It has also proactive component (a fifth “P”) instead of being reactive.

The idea then is that subjacent to the human morphological and physiological individuality, there is also a genomic individuality. All the physical, intellectual, and behavioral characteristics of individuals at any given time are determined by both their genome and their life history. We thus have the genomic paradigm of health as the harmonious balance between genome and environment. The corollary of this is that disease represents the genome/environment disequilibrium.

Precision Medicine emerges naturally from this genomic paradigm of health and disease. Knowing the intimacy of the genomic variations that determine predispositions and resistance, it is possible to manipulate the environment (lifestyle, diet, addition or removal of drugs, preventive surgery, frequency of clinical and laboratory tests) in order to maintain the harmonious genome/environment balance that characterizes health.

Traditionally there were two major aspects of medicine. One was Curative Medicine, of a personal nature, treating highly symptomatic (that is, already sick) patients with low efficiency, since few human diseases can be effectively cured. The second was Preventive Medicine, aimed at maintaining public health. Precision Medicine brings together the best of these two strands, allowing the practice of a medicine that is both preventive and personalized. It aims at caring for patients still asymptomatic or barely symptomatic, with high efficiency, in order to prevent or retard the development of diseases. Genomic Medicine did not come to replace the pre-existing medical aspects, but to add to them.

The practice of Precision Medicine does not differ substantially from the traditional clinical medicine that doctors have incorporated over the years, into their decisions on routine laboratory and image exams. In the twenty-first century, we are similarly incorporating the testing of genomic individuality, which can now be done efficiently using sequencing techniques that allow us to study individual variations in hundreds of thousands of genes simultaneously, at a lower cost each day. The results of the tests allow us to practice what was previously impossible: truly personalized medicine, backed by the intimate knowledge of the patient's own genetic makeup.

The concept of such genetic check-up can be conceptualized as the same as that of a battery of multiple laboratory tests, each with the power to edit different phenotypes. Each DNA variant discovered is comparable to an individual medical test, which provides the doctor with a post-test probability that a given particular unwanted phenotype will occur. This opens up the possibility of modulating the environment to adapt it to the genotype of the individual, aborting the genesis of the disease. Such is the essence of Precision Medicine.

I would like to recall here, an aphorism of the great physician William Osler (1849–1919), who is considered the father of scientific medicine and who prophetically incorporated into his teachings many of the aspects of Precision Medicine: "If it were not for the great variability among individuals, medicine might as well be a science and not an art". Imbedded into this, is the notion that if we can achieve success in fully understanding and characterizing the variability among individuals, we will be able to rescue medicine as a science.

I wish to discuss how we can implement in the broadest possible sense this Precision Medicine in practice. In particular, I ask the questions: (1) Is whole genome sequencing ripe for routine utilization in medicine? (2) Should we sequence everybody's genomes? We all know that the use of whole genome sequencing (WGS) and whole exome sequencing (WES) have afforded us with hereto unthinkable diagnostic powers to diagnose Mendelian diseases, especially rare monogenic diseases (Fig. 1.3). However, most common human diseases (breast cancer, ovarian cancer, colorectal cancer, coronary artery disease, type 2 diabetes, autoimmune diseases, psychoses, etc.) are not predominantly Mendelian and monogenic, but instead are caused by numerous common predisposing variants ("polygenic") in interaction

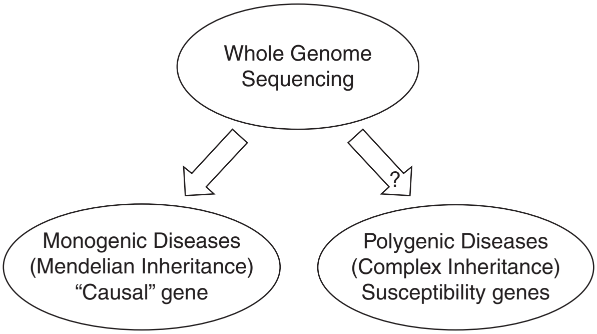

Whole Genome Sequencing
?
Monogenic Diseases (Mendelian Inheritance) "Causal" gene
Polygenic Diseases (Complex Inheritance) Susceptibility genes

Fig. 1.3 Whole genome sequencing has shown its true colors in the diagnosis of monogenic Mendelian diseases, but it is questionable if it will be equally useful in accurately assessing the polygenic component of inherited risk

with the environment. Comprehensive genome interpretation should enable assessment for both monogenic and polygenic components of inherited risk (Fig. 1.3). Can we use WGS to achieve both?

1.4.2 New-Generation DNA Sequencing (NGS)

The traditional method of DNA sequencing was developed in Cambridge, England, by the scientist Fred Sanger, who won a Nobel Prize (his second) for this important advance. The method, based on the use of dideoxynucleotides, is colloquially called Sanger Sequencing. All the Human Genome Project was executed using this methodology, which still stands as the gold standard for sequencing in terms of accuracy and reliability, especially since it allows analyses of DNA fragments several hundred base pairs long. The method is slow, as it depends on an electrophoresis stage, which today is performed by capillary electrophoresis in automatic fluorescent sequencers. Presently, Sanger Sequencing is mainly limited to the analysis of individual genes and to confirm some variants found in New Generation Sequencing.

New Generation Sequencing (NGS; also called Maximal Parallel Sequencing) was developed after the end of the Human Genome Project and represented a huge revolution in genomic analysis. NGS allows, in a single sequencer, a sequencing speed of 100,000–200,000 times higher than that with the Sanger method. It allows the sequencing of an entire human genome in less than a week.

In NGS, there is no need to purify a fragment to be sequenced. Millions of DNA fragments present in a complex mixture that can contain tens, hundreds or thousands of genes, are sequenced simultaneously, in parallel, without the need for electrophoresis. Because of this enormous speed, the price of sequencing has fallen off considerably, being now possible to sequence a whole human genome for less than US $ 1000 in reagents.

The main disadvantage of NGS is that a gene is not sequenced continuously, but in thousands of small fragments of 100–150 base pairs, which then need to be correctly concatenated by comparison with a reference sequence. Since the genome sequence is aligned by bioinformatic annealing to the reference sequence, formally speaking, this is not true de novo sequencing, like that done in the Human Genome Project, but rather resequencing. For instance, it assumes that the general organization of the genome in copy number of specific fragments is the same as in the reference genome. We know that this is not true because there is variation in structural features between different people [30]. Also, by sequencing only small fragments, in general, NGS is not able to diagnose mutations created by large insertions (e.g. Alu sequences) or microsatellite expansions. However, this and other difficulties are likely to be overcome in the near future as new techniques are optimized for long-range sequencing or fast de novo genome assemblage [31].

1.4.3 Pathogenic Mendelian Variants ("Monogenic")

Mendelian diseases affect at least one person in every 50. We know about 4500 Mendelian diseases, but estimates are that there may be at least twice as many [32]. Although individually rare, together they generate a large burden for public health. In developed countries, it is estimated that about 50% of patients with a rare Mendelian disease are never properly diagnosed. Arriving at an accurate molecular diagnosis in a Mendelian disease has a number of advantages:

- Puts an end to the diagnostic odyssey.
- Improves the quality of medical monitoring of the disease, including possible treatments, establishment of prognoses and prevention of complications.
- Allows genetic counseling of families, regarding the risk of recurrence, prenatal diagnosis options and pre-implantation diagnosis.
- It allows the exorcism of parental beliefs and erroneous hypotheses about the cause of the disease.
- It allows emotional closure by parents.

Whole Genome Sequencing (WGS) and its less expensive sib Complete Exome Sequencing (WES) have come to try to resolve cases that remain undiagnosed after detailed and intensive investigations. The evidence in the literature is that WGS when performed on patients and their parents (three genome sequences) allows the definitive diagnosis and identification of the genetic defect in 30–50% of the patients evaluated for suspected genetic disease.

Anyway, a fundamental element is who analyzes the variants found in the sequencing; ideally a professional who has clinical experience in both medical genetics and bioinformatics expertise. Thus, the same professional can make the best possible assessment of the pathogenicity of the variant(s) found and integrate the results with the clinical picture to arrive at the correct diagnosis.

Among geneticists, there has been a heated debate about the use of WGS versus WES for the diagnosis of genetic diseases. As the name implies, WGS seeks to sequence the entire genome. Due to the difficulty in sequencing technically challenging regions of the genome with current sequencing platforms (regions with high GC content, repeatable regions, centromeres, telomeres, etc.), WGS covers only about 95% of the genome, although it sequences more than 99.7% of the exons. WES is not capable of finding variants when they are not in an exon (15% of the mutations observed in Mendelian diseases are not in an exon). WGS is able to overcome some of these limitations—it is able to diagnose variants in promoter regions, in other regulatory regions (enhancers) and in the middle of introns, although it still is necessary for the specific variant to have already been described previously. In fact, the ability to diagnose unknown variants in the noncoding portion of the genome to identify regulatory mutations is still limited.

WES depends on a step of exon capture and has the added limitation that it does not detect pathogenic variants when they are in an exon selected in low efficiency by the capture technique. This occurs especially in regions rich in GC [33]. WGS does not have this problem, since it does not depend on a capture step. Thus, WGS provides even coverage and is more powerful than the WES for detecting single nucleotide variants and indels even in areas well covered by the capture kit [34]. In addition, WGS is capable of detecting more CNVs as covers all break points and detects variants in protein and RNA coding regions that are outside the capture kit's coverage. In other words, WGS is much more powerful than WES for diagnosis.

Currently, WGS costs roughly twice as much as WES. However, most of the cost of WGS comes from the sequencing itself, while the cost of WES is mainly due to the price of the capture kit. As the costs of sequencing continue to fall, while the price of the capture kit remains more or less stable, there will be a time when the cost of WGS will come close to WES. Thus, it is likely that in the near future the WGS may replace WES in the analysis of human genetic diseases.

But the question that we wish to discuss here is: can we use WGS of healthy individuals to acquire knowledge of our patients' genomes that will permit us to put into practice the Precision Medicine or P4 Medicine and impact positively their medical future?

The conventional criteria for evaluating genetic tests include analytic validity, clinical validity, and clinical utility. In sequencing, analytical validity refers to a test's ability to measure the genotype of interest accurately and reliably. Clinical validity refers to a test's ability to detect or predict the clinical disorder or phenotype associated with the genotype. On the other hand, clinical utility is a measure of its usefulness in the clinic and resulting changes in clinical endpoints. Clinical validity is predicated on the assumption that there is a scientifically valid association between the gene and trait. Thus, scientific validity is a prerequisite for clinical validity, but not the only component. Clinical validity also encompasses the predictive value of the test, which can be called predictive ability [35, 36].

To misquote Shakespeare, therein lies the rub! The study of whole genome sequencing in a healthy individual provides information that is purely genotypic. Even if a well-known pathogenic variant is found on WGS, we cannot predict that the variant will cause a disease phenotype because of the problem of penetrance. We can define penetrance of a single gene disorder as the probability that a person who has a pathogenic variant will express the disease phenotype [4]. The opposite of full penetrance is incomplete penetrance or non-penetrance.

We can understand better the problem by using a true case as an example. In 2019, I agreed to perform in our laboratory the WGS of a healthy journalist who wanted to write an article about the process of sequencing his own genome. In the analysis of the genome, I encountered the variant NM_000552.3(VWF):c.2561G > A p.(Arg854Gln) which is listed in the ClinVar Databank (https://www.ncbi.nlm.nih.gov/clinvar) as "pathogenic" for types 1 e 2 of Von Willebrand Disease, a coagulation disorder (RCV000507204.1, RCV000169683.3, RCV000086620.3, RCV000336497.1, RCV000000321.2 e RCV000000321.2). I suggested that he should consult with a hematologist and undergo the appropriate tests. As he recounted in his article [37] the assay for the Von Willebrand factor in his blood was 85.5%, well within the normal range (50–160%). That was easily solved because there was an efficient and easily accessible test for Von Willebrand disease. What if instead of identifying a pathogenic variant for Von Willebrand, I had identified a pathogenic variant for a severe neurological disease with appearance only at advanced age?

Reduced penetrance of well characterized pathogenic variants for autosomal dominant diseases have been described in a myriad of diseases, including cardiac arrythmia syndromes [38], hypertrophic cardiomyopathies [39], immunodeficiencies [40] and cancer-susceptibility mutations [41], just to mention a few. Moreover, reduced penetrance may explain not only why genetic diseases are occasionally transmitted through unaffected parents, but also why healthy individuals can harbor quite large numbers of potentially disadvantageous variants in their genomes without suffering any obvious ill effects [42].

Even if we disregard the more obvious and trivial "age-related non-penetrance" for late appearing diseases and "sex-related non-penetrance" for sex-specific diseases, reduced penetrance is still very widespread. After all, Mendelian diseases involve a single gene, while there are more than 20,000 in the human genome, with plenty of opportunity for genetic or epigenetic modification of the phenotypic effects of the pathogenic mutant allele. Environmental factors may also play a part. We cannot go in great detail in here and the reader in search of more detailed information should access the very complete review by Cooper et al. [42], from which I obtained the inspiration for Fig. 1.4.

One upon a time, it was believed that, at least for "monogenic" disorders, genotype–phenotype relationships would be simple and easy to establish. However, now it does not seem appropriate to regard such disorders as either simple or *sensu strictu* monogenic [42]. In other words, reduced penetrance possibly occurs for every disease, arising from the complex interplay between the overabundance of genetic variation present in the human genome and environmental factors [40]. Certainly, the pathogenic variant may be monogenic, but its disease expression is multifactorial, probably often with a polygenic component.

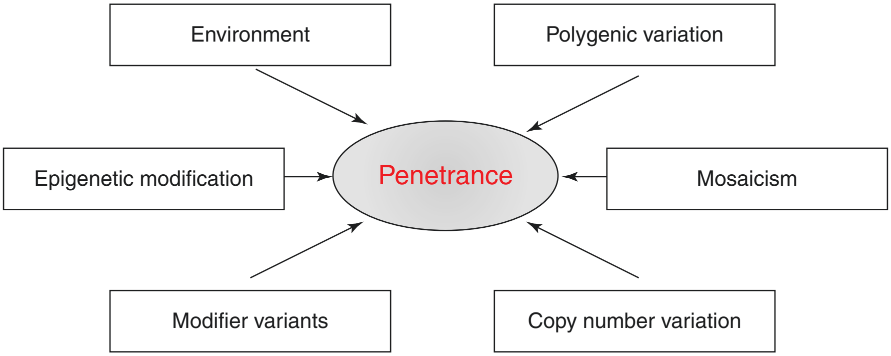


Fig. 1.4 Some factors that may modulate the clinical penetrance of a pathogenic Mendelian variant (modified from [42])

1.4.4 Cumulative Effect of Numerous Common Variants ("Polygenic")

For the past 15 years, genome-wide association studies (GWAS) have contributed to the identification of the association of specific genomic regions with an impressive number of common diseases, including breast cancer, ovarian cancer, colorectal cancer, coronary artery disease, type 2 diabetes, autoimmune diseases, psychoses, etc. By September 2018, the NHGRI-EBI catalog of such studies contained 5687 GWAS comprising 71,673 variant-trait associations from 3567 publications [43].

Polygenic Risk Score (PRS) is a single risk figure that is meant to incorporate the aggregate effect of thousands of genetic variants across the human genome into a single score [44]. For that, one uses the sums of the effects of all single nucleotide polymorphisms (SNP) typed, weighted by the magnitude of the association between the genotype at a particular SNP and the trait of interest. There are multiple approaches for constructing polygenic risk scores, ranging from including only of SNPs that have exceeded genome-wide significance thresholds to the more modern use of millions of SNPs encompassing those that individually only very weakly associate with the phenotype of interest.

More recently there have been proposals that the "standard" method of calculating Polygenic Risk Scores by genome-wide association studies (GWAS) followed by imputation, can be profitably replaced by WGS [45]. Indeed, Khera et al. [46] have shown in detail how to obtain a Polygenic Risk Score using only new-generation sequencing data.

For clinicians, the promise of the method of Polygenic Risk Scores for estimation of the risk of a disease with complex inheritance (polygenic disease) is very appealing. In consequence, this has led, as wittily put by Rotter and Lin [47], to "an outbreak of polygenic scores for coronary artery disease". The same has happened to numerable other diseases with complex inheritance, as can be easily ascertained by a perusal of Pubmed.

Nonetheless, serious doubts about the value of Polygenic Risk Scores have emerged both from experimental and theoretical perspectives. Experimentally, some careful studies have failed to confirm the clinical utility and clinical validity of Polygenic Risk Scores. For instance, a recent retrospective cohort study assessed 7237 middle-aged participants of European ancestry free of clinical coronary heart disease at baseline. When they added a polygenic risk score to the 2013 American College of Cardiology and American Heart Association pooled cohort equations, it did not significantly improve discrimination, calibration, or risk reclassification compared with conventional predictors. They concluded that a polygenic risk score may not be able to enhance risk prediction in a general, white middle-aged population (Mosley et al., 2020).

In my opinion, the most severe criticism came on theoretical grounds. In a very cogent recent article, Wald and Old [49] observed that hopes that individuals identified at by high polygenic risk scores might benefit from preventive interventions rest on the incorrect assumption that the odds ratios derived from polygenic risk scores are also directly useful in risk prediction and population screening. The authors point out that estimates of the relative risk between a disease marker and a disease have to be extremely high for the risk factor to merit consideration as a worthwhile screening test. According to them, we should avoid unrealistic expectations in medical screening [49]. The most prudent attitude at the moment should be conservative. We should avoid the hype and hold any medical use of Polygenic Risk Scores until further studies and publications have definitively established their clinical utility and clinical validity.

In conclusion, review of available evidence does not favor the idea that at this moment in time whole genome sequencing is sufficiently developed to allow reliable predictions of monogenic and polygenic components of inherited risk in healthy individuals. WGS should still be reserved for the diagnosis of pathogenic variants of Mendelian diseases.

References

1. Goldman AD, Landweber LF. What is a Genome? PLoS Genet. 2016;12(7):e1006181.
2. GHR – Genetics Home Reference (2020). What is a genome? https://ghr.nlm.nih.gov/primer/hgp/genome. Accessed 29 Aug 2020.
3. Brown TA. Genomes 4. New York: Garland Science; 2017.
4. Strachan T, Goodship JC, Hinnery P. Genetics and genomics in medicine. New York: Garland Science; 2015.
5. NHS National Genetics and Genomics Education Centre. Diagram of components of the genome as estimated in 2014. https://commons.wikimedia.org/w/index.php?curid=50582882. Accessed 29 Aug 2020.
6. Graur D, Li W-H. Fundamentals of molecular evolution. 2nd ed. New York: Sinauer; 2000.
7. Varki A, Geschwind DH, Eichler EE. Explaining human uniqueness: genome interactions with environment, behaviour and culture. Nat Rev Genet. 2008;9(10):749–63.
8. Graur D, Zheng Y, Price N, et al. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5(3):578–90.
9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
10. Lee H, Zhang Z, Krause HM. Long noncoding RNAs and repetitive elements: junk or intimate evolutionary partners? Trends Genet. 2019;35(12):892–902.
11. Koonin EV. Darwinian evolution in the light of genomics. Nucleic Acids Res. 2009;37(4):1011–34.
12. Orlean S. The library book. New York: Simon & Schuster; 2019.
13. Borges J. Labyrinths. New York: Book-of-the-Month Club; 1962.
14. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10(10):691–703.
15. Richardson SR, Doucet AJ, Kopera HC, et al. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol Spectr. 2015;3(2):MDNA3-2014.
16. Matheson NJ, Lehner PJ. How does SARS-CoV-2 cause COVID-19? Science. 2020;369(6503):510–1.
17. Johnson WE. Endogenous retroviruses in the genomics era. Annu Rev Virol. 2015;2(1):135–59.
18. Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. Nat Rev Microbiol. 2019;17(6):355–70.

19. Koonin EV, Dolja VV. A virocentric perspective on the evolution of life. Curr Opin Virol. 2013;3(5):546–57.
20. Zhang YZ, Wu WC, Shi M, Holmes EC. The diversity, evolution and origins of vertebrate RNA viruses. Curr Opin Virol. 2018;31:9–16.
21. Hayward A. Origin of the retroviruses: when, where, and how? Curr Opin Virol. 2017;25:23–7.
22. Grandi N, Tramontano E. Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses. Front Immunol. 2018;9:2039.
23. Katsura Y, Asai S. Evolutionary medicine of retroviruses in the human genome. Am J Med Sci. 2019;358(6):384–8.
24. Luganini A, Gribaudo G. Retroviruses of the human virobiota: the recycling of viral genes and the resulting advantages for human hosts during evolution. Front Microbiol. 2020;11:1140.
25. Hurst TP, Magiorkinis G. Activation of the innate immune response by endogenous retroviruses. J Gen Virol. 2015;96(Pt 6):1207–18.
26. Gould SJ, Vrba ES. Exaptation - a missing term in the science of form. Paleobiology. 1982;8:4–15.
27. Monde K, Terasawa H, Nakano Y, et al. Molecular mechanisms by which HERV-K Gag interferes with HIV-1 Gag assembly and particle infectivity. Retrovirology. 2017;323(7):627–35.
28. Garcia-Montojo M, Doucet-O'Hare T, Henderson L, Nath A. Human endogenous retrovirus-K (HML-2): a comprehensive review. Crit Rev Microbiol. 2018;44(6):715–38.
29. Obama B. The precision medicine initiative. 2015. https://obamawhitehouse.archives.gov/precision-medicine. Accessed 29 AUG 2020.
30. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172–83.
31. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. Front Genet. 2019;10:426.
32. Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. Am J Hum Genet. 2019;105(3):448–55.
33. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? Hum Genet. 2016;135(3):359–62.
34. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015;112(17):5473–8.
35. Grosse SD, Kalman L, Khoury MJ. Evaluation of the validity and utility of genetic testing for rare diseases. Adv Exp Med Biol. 2010;686:115–31.
36. McCarthy J. How to determine the clinical validity of genetic tests. Precision Medical Advisors. 2019. https://www.precisionmedicineadvisors.com/precisionmedicine-blog/2019/8/2/how-to-evaluate-health-related-genetic-tests. Accessed 29 Aug 2020.
37. Leite M. My dear genome. Folha de São Paulo, Ilustríssima. (03/19):4–5; 2019.
38. Giudicessi JR, Ackerman MJ. Determinants of incomplete penetrance and variable expressivity in heritable cardiac arrhythmia syndromes. Transl Res. 2013;161(1):1–14.
39. Sabater-Molina M, Pérez-Sánchez I, Hernández Del Rincón JP, Gimeno JR. Genetics of hypertrophic cardiomyopathy: a review of current state. Clin Genet. 2018;93(1):3–14.
40. Gruber C, Bogunovic D. Incomplete penetrance in primary immunodeficiency: a skeleton in the closet. Hum Genet. 2020;139(6–7):745–57.
41. Tung N, Domchek SM, Stadler Z, et al. Counselling framework for moderate-penetrance cancer-susceptibility mutations. Nat Rev Clin Oncol. 2016;13(9):581–8.
42. Cooper DN, Krawczak M, Polychronakos C, et al. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013;132(10):1077–130.
43. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–12.
44. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019;28(R2):R133–42.

45. Homburger JR, Neben CL, Mishne G, et al. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. Genome Med. 2019;11(1):74.
46. Khera AV, Chaffin M, Zekavat SM, et al. Whole-Genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. Circulation. 2019;139(13):1593–602.
47. Rotter JI, Lin HJ. An outbreak of polygenic scores for coronary artery disease. J Am Coll Cardiol. 2020;75(22):2781–4.
48. Jonathan D. Mosley, Deepak K. Gupta, Jingyi Tan, Jie Yao, Quinn S. Wells, Christian M. Shaffer, Suman Kundu, Cassianne Robinson-Cohen, Bruce M. Psaty, Stephen S. Rich, Wendy S. Post, Xiuqing Guo, Jerome I Rotter, Dan M. Roden, Robert E. Gerszten, Thomas J. Wang. Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. JAMA. 2020;323(7):627.
49. Wald NJ, Old R. The illusion of polygenic disease risk prediction. Genet Med. 2019;21(8):1705–7.

Chapter 2
Human Chromosomes

Bianca Pereira Favilla, Luciana Amaral Haddad, and Maria Isabel Melaragno

2.1 Introduction

The double-helix model was proposed as a molecular structure for the deoxyribonucleic acid (DNA) by James Watson and Francis Crick in 1953 [1], based on DNA chromatographic data from Erwin Chargaff [2] and X-ray and diffraction analyses by Maurice Wilkins [3] and Rosalind Franklin [4, 5]. The DNA molecule is the association of two antiparallel, complementary polynucleotide chains in a right-handed helix. In DNA, each polynucleotide chain is composed of deoxyribonucleosides ligated by phosphodiester bond. The nucleoside is a pentose sugar, which in the case of DNA is a deoxyribose, having its 1′-carbon associated by a glycosidic bond to a nitrogenous base either adenine (A), cytosine (C), guanine (G) or thymine (T). The pentose carbons are numbered 1′ to 5′ to differentiate from the nitrogenous base carbons. DNA nitrogenous bases classify as purines (adenine or guanine) and pyrimidines (cytosine or thymine). Phosphorylation of deoxyribonucleosides leads to the formation of deoxyribonucleotides, such as deoxyribonucleoside mono-, di- or triphosphates. The phosphodiester bond forms upon the attack by the 3′-hydroxyl from a nucleotide towards the deoxyribonucleoside 5′-triphosphate, leading to hydrolysis between the α- and β-phosphates and release of pyrophosphate. Consequently, the hydroxyl at the 3′ end of the nucleotide becomes esterified to the alpha phosphate of the reacting nucleotide (Fig. 2.1).

B. P. Favilla · M. I. Melaragno (✉)
Genetics Division, Department of Morphology and Genetics, Universidade Federal de São Paulo, São Paulo, Brazil
e-mail: bfavilla@unifesp.br; melaragno.maria@unifesp.br

L. A. Haddad
Department of Genetics and Evolutionary Biology, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil
e-mail: haddadL@usp.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_2

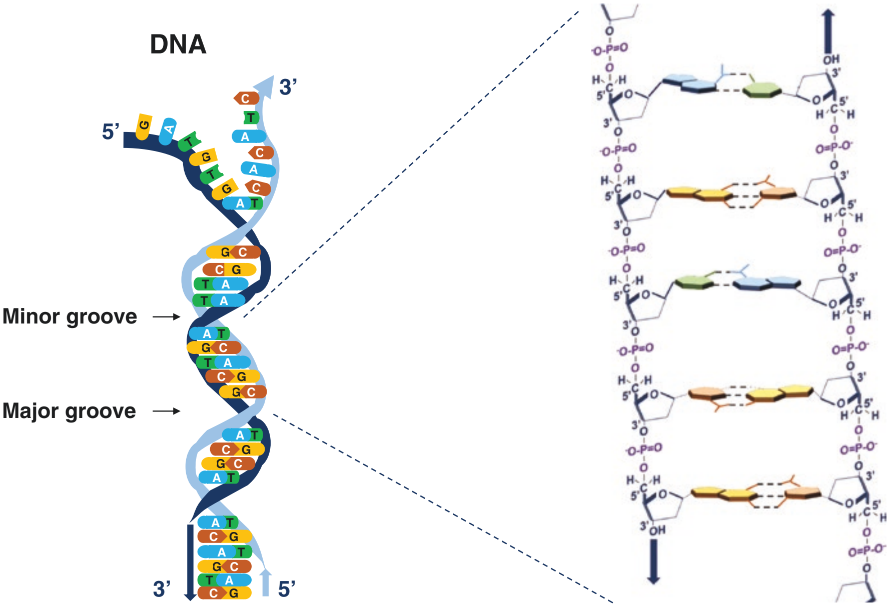


Fig. 2.1 Diagram of a short segment of a DNA molecule, showing on the left the double helix with its anti-parallelism (5′ and 3′ ends identified) and the upper side denatured. Arrows show the 5′-to-3′ polarity that is the direction of chain growth during DNA replication. A 5-bp segment of the molecule is enlarged on the right side of the figure, depicting the phosphorylated 5′ end and the reacting hydroxyl at the 3′ carbon of the pentose. Base pairing is indicated between A and T or C and G, respectively by two and three hydrogen bonds

The two chains of DNA, each one known as a DNA strand, interact by hydrogen bonding between puric and pyrimidic nitrogenous bases, allowing for base pairing between adenine and thymine (two hydrogen bonds) or guanine and cytosine (three hydrogen bonds). This is in agreement with Chargaff’s chromatographic findings of nearly equal amounts of purine and pyrimidine bases in all living beings examined, as well as with Wilkins and Franklin’s estimation of a constant external diameter of 2 nm and internal distance of 1.1 nm for the DNA helicoidal structure.

The antiparallel characteristic of the DNA double helix refers to the opposite polarities of the two chains. While one chain is 3′ to 5′, *i.e.*, oriented from the sugar 3′ carbon of one nucleotide on one end towards the sugar 5′ carbon of the succeeding nucleotide, the complementary chain is in the opposite orientation, which is 5′ to 3′. Every ten base pairs (bp), the helix completes a 360-degree trajectory corresponding to 3.4 nm in length. Thus, two consecutive base pairs are less than 4 Angstrom apart, disposed in a parallel manner (Fig. 2.1).

Considering the DNA molecule as a double helix with a diameter of two nanometers, its ultrastructural external surface is remarkable for two longitudinal grooves running helicoidally along the cylindrical shape. The wider groove, named major groove, consists of the space between adjacent gyri. The narrowest groove, called minor groove, is the space between the sugar phosphate backbones of complementary strands (Fig. 2.1). The major and minor grooves are important sites for protein interactions with the DNA.

2.2 DNA Replication

The replication of DNA occurs in the S (synthesis) phase of the cell cycle (Fig. 2.2), and is semi-conservative as supported by evidences by Meselson and Stahl [6]. Each strand from the parental molecule will be fully conserved in a daughter molecule,

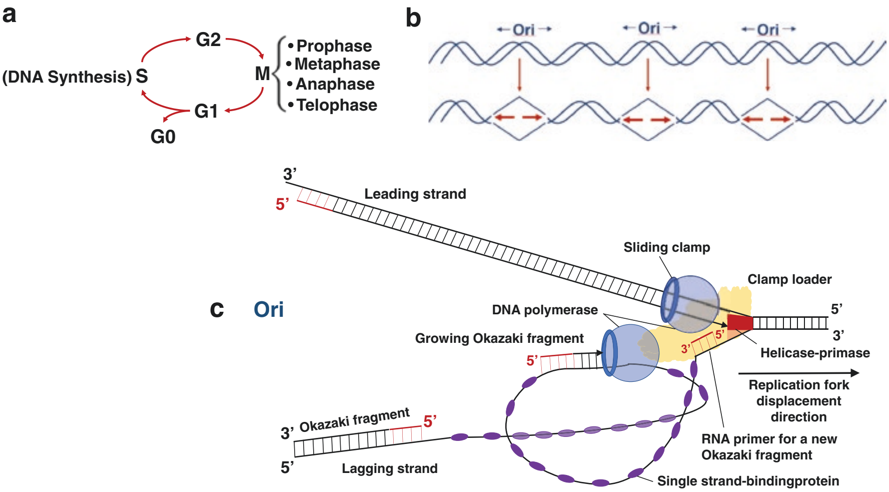


Fig. 2.2 (**a**) DNA replication occurs in the S (synthesis) phase of the cell cycle, before the mitosis phase that allows for segregation of chromatids and is concluded upon cell division. The four stages of the M phase are indicated. The S and M phases are separated by gap or growth phases (G1 and G2), which are periods marked by gene expression, anabolism and consequently pronounced cell growth. Cell differentiation may occur at G2. Cells that enter senescence (G0) exits the cell cycle. (**b**) A linear chromosome has many origins of DNA replication (ori) where DNA starts denaturation forming two replication forks (left and right horizontal arrows) that will lead to further denaturation of DNA on each side of ori. (**c**) One replication fork from one ori is depicted. The same process should take place on the replication fork on the other side of ori, except for the opposite strand polarities. The leading and lagging strands are indicated with their polarities. On the replication fork, helicase and primase stand associated with the clamp loader, which is responsible to load and maintain one DNA polymerase and its sliding clamp on the leading strand, and one DNA polymerase and a sliding clamp for every Okazaki fragment under synthesis on the lagging strand. Although only one polymerase is illustrated for the lagging strand, it is believed that on the lagging strand two polymerases may be at work at a time, simultaneously synthetizing two Okazaki fragments, though at different polymerization stages. The sliding clamp increases the processivity of the polymerase. RNA primers are in red. A fully formed Okazaki fragment is presented on the left. Single strand-binding proteins (purple) stabilize DNA single strands during the replication process

and serves as template for the synthesis of a new, complementary strand. Therefore, upon DNA replication, each of the two daughter DNA molecules produced consists of a parental strand and a newly synthetized strand. The semi-conservative replication of DNA allows for copying the full set of DNA sequence of each chromosome.

The DNA polymerase catalyzes the formation of a phosphodiester bond between the 5′-triphosphated end of a free nucleoside triphosphate and the hydroxyl at the 3′ end of a nucleotide in a growing polynucleotide chain hybridized to the DNA template strand. The incoming nucleotide triphosphate needs to be complementary to the next position on the template. Once the phosphodiester bond formation is catalyzed and the pyrophosphate released, the polymerase shifts to the next position on the DNA template. However, a new reaction will only take place if the recently incorporated nucleotide is correctly paired with the template. This proofreading is performed by the DNA polymerase itself, which will correct the mispaired nucleotide, if detected. The correction is allowed by the polymerase own 3′-to-5′ exonuclease function, hydrolyzing the recently formed phosphodiester bond and starting over. The DNA polymerase proofreading activity and its 3′-to-5′ exonuclease ability allow for 10^3- to 10^4-fold reduction of made by the DNA polymerase during DNA replication, thus significantly contributing to decrease the fixation of sequence errors that would originate mutations in the DNA.

In order to provide two templates for DNA replication, the two strands of the parental DNA molecule need to denature from a starting point denominated origin of replication (Fig. 2.2). As eukaryotes have long, linear chromosomes, each of them has many origins of replication. In human chromosomes, origins of replication are assumed to be nearly 40 kilobases (kb) apart. Once the DNA is denatured by the action of helicases on the origin of replication, two replication forks are formed and each one will proceed denaturing the DNA in opposite directions (Fig. 2.2). Since DNA polymerases can only add nucleotides to the 3′ end of the nascent strand, the DNA synthesis takes place from 5′ to 3′ in a growing polynucleotide chain that is antiparallel to the template. This particularity of the DNA polymerase action poses a challenge to replicate both DNA strands at the same time on a fork that shifts in a single direction. On the DNA template strand oriented from 3′ to 5′ along the direction of the replication fork, the new strand will grow continuously from 5′ to 3′, following the progression of the replication fork. This strand of continuous DNA replication is known as the leading strand. On the other hand, the DNA template strand oriented from 5′ to 3′ along the direction of the replication fork will have the same orientation expected for the DNA synthesis to take place. Therefore, DNA synthesis will need to occur on the direction opposite to the shifting of the replication fork. To circumvent this problem, the DNA template strand on the 5′ to 3′ orientation displaces into a loop near the fork, and the synthesis of the new strand must take place discontinuously by approximately 150-nucleotide fragments, known as Okazaki fragments after their discovery by Reiji and Tsuneko Okazaki [7]. The strand of discontinuous DNA replication is known as the lagging strand. On each fork established from an origin of replication one strand will grow continuously and the other discontinuously. Therefore, the DNA replication is said to be semi-discontinuous (Fig. 2.2).

As presented above, the DNA polymerase catalyzes the formation of a phosphodiester bond between an incoming nucleotide and the 3′ end of a growing chain, but it is not able to start DNA polymerization without a primer or a template. Since it does not polymerize *de novo*, it must have a primer to initiate the DNA replication. RNA primers synthetized by an enzyme known as primase are thus essential for DNA replication. There is one RNA primer to initiate DNA synthesis on the leading strand and one primer for every Okazaki fragment (Fig. 2.2). As primase has high affinity for the helicase, this protein heterodimer acts on the fork, and the two enzymes are simultaneously activated. Upon helicase activation, a further stretch of DNA is denatured and the primase synthesizes a new RNA primer for another Okazaki fragment. The RNA primers are short, and need to be excised before the cell exits the S phase. RNA primers and Okazaki fragments are repaired in the same way. As the primers have free 5′ ends unattached to other fragments, ribonucleotides from the primers are hydrolyzed by the 5′ to 3′ exonuclease activity of the DNA polymerase or by the RNAse H that acts on hybrids of DNA and RNA. The gaps produced due to primer hydrolysis are filled in by the DNA polymerase, and then full DNA segments are ligated by the DNA ligase. The repair of the 3′ overhang on chromosome telomeres after DNA replication is presented in Chap. 7.

2.3 Chromatin

Eukaryotic chromosomes are long DNA molecules with hundreds of genes, a significant part of which must be transcriptionally repressed in certain cell types and developmental stages. Considering the distance between two base pairs as approximately 0.34 nm, the length of the human chromosome-1 DNA, harboring nearly 250 Mbp, should be around 8.5 cm. The organization of the DNA into chromatin has enabled the chromosomal DNA to fit into the eukaryotic cell nucleus (average length of 15μm) and to limit the access of general transcription factors to gene promoters (see Chap. 4).

The nucleosome is the basic unit of the chromatin, and consists of a disc-shaped core of eight histone proteins wrapped by a segment of 147-bp DNA in a left-handed way and 1.7 helical turn (Fig. 2.3). The length of 147 bp of DNA associated with the histone core is an invariable characteristic of the nucleosome independently on the eukaryotic species. The histone octamer is composed of two units of each of the four histones H2A, H2B, H3, and H4. The nucleosome assembly starts with the binding of the H3-H4 tetramer to the DNA, and follows with the association of two H2A and H2B heterodimers. The DNA between two nucleosomes, known as linker DNA, is devoid of histones and has a variable length among different species. In the human chromatin, the linker DNA has in average 40 to 50 base pairs.

Histones H2A, H2B, H3 and H4 have low molecular masses (11 to 15 kDa). A characteristic of the histone protein family is the high content (at least 20%) of basic (positively charged) residues that allow for electrostatic interactions with the negative charge of DNA phosphates. In addition, hydrogen bonding between histones

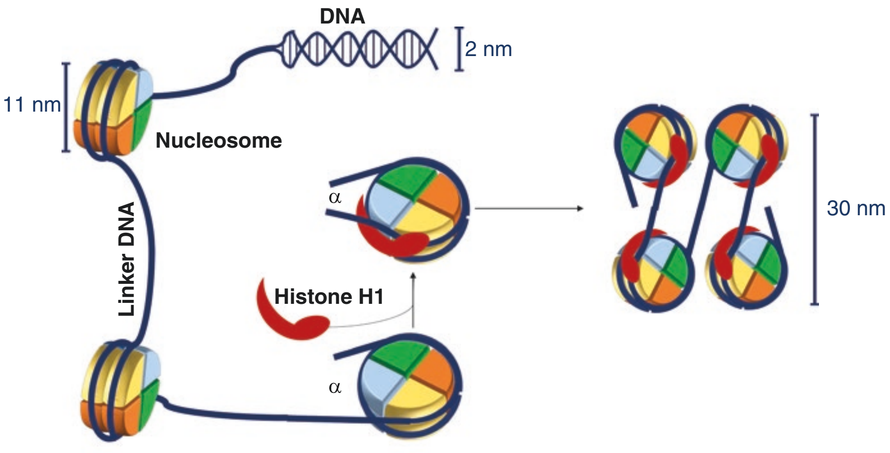


Fig. 2.3 Chromatin formation. The nucleosome formation is shown leading to the assembly of the 11-nm chromatin fiber by the association of a histone octamer and the DNA wrapping it in a nearly 1.7 turn. The linker DNA is shown between two nucleosomes. The angle (α) formed between two linker DNA segments from the same nucleosome is wider on the 11-nm fiber than in the 30-nm fiber, which is defined upon the binding of histone H1 to the nucleosome. This narrowing makes nucleosomes come closer to each other in a zig-zag fashion that is the basis for the assembly of the 30-nm chromatin fiber, and its different proposed models (not shown in this figure)

and oxygens from the DNA sugar phosphate backbone within the minor groove helps to stabilize the nucleosome. Therefore, the nucleosome formation does not depend on the DNA sequence.

The initial structural observation of the chromosomes at the optical microscope considered that chromosomes co-existed in the same cell nucleus as heterogeneous chromogenic material, ranging in various degrees from the lightest coloration and least condensation (euchromatin) to the darkest coloration and highest condensation (heterochromatin). The microscopic and biochemical assay data had indicated that the heterochromatin could consist of a higher order of organization of DNA and histone units. The 1970's works by Kornberg and Thomas [8] and Olins and Olins [9] on the characterization of the nucleosome were fundamental for the proposition of the chromatin as a repetitive structure of the histone octamer and a nearly 200-bp DNA segment [10].

Along the cell cycle and according to the gene expression activity (see Chap. 4), chromatin may be observed on the electron microscope at distinct condensation levels that reflect the degrees of the DNA compaction. The first level of DNA compaction is provided by the formation of the nucleosome, the diameter of which (11 nm) has named this configuration as the 11-nm fiber by contrast to the 2-nm diameter of the DNA double helix (Fig. 2.3). On the electron microscope, the 11-nm fiber can be observed in a morphological chromatin configuration like beads in a string.

Upon the binding of a fifth member of the histone protein family, histone H1 (21 kDa), to the nucleosome and the linker DNA on one side, the 11-nm chromatin organization level becomes susceptible to a more organized and compact level, known as the chromatin 30-nm fiber or the second level of DNA compaction.

Histone H1 binding allows for a further 20-bp wrapping of DNA around the histone octamer. It tightens the nucleosomal structure and defines a lesser angle between linker DNA segments adjacent to a nucleosome (Fig. 2.3). Consequently, consecutive nucleosomes will be closer to each other and will tend to position in a more organized, compact, zig-zag fashion. However, this difference in DNA compaction is not yet sufficient for the formation of the 30-nm fiber.

The structure and assembly of the 30-nm fiber are still a matter of debate. Among different models proposed for the assembly of the 30-nm fiber, the solenoid model has been considered for more than 40 years [11]. The solenoid model predicts a supercoiled structure of the chromatin, in which every turn would have six to eight nucleosomes in a radial disposition, with the planar surface of the histone disc directed to the following one in an angle of nearly 36°. Although the histone H1 is at the center of the super-helix in the solenoid model, it does not cross the axis. By contrast, in the zig-zag-based coil models, the DNA linker crosses the central axis, which is more consistent with longer DNA linkers; and the nucleosomes' cores stack helically aside. For any model of the chromatin 30-nm fiber, the addition of histone H1 is necessary for the assembly of the 30-nm super-helix of nucleosomes (Fig. 2.3).

Increasing levels of chromatin condensation and DNA compaction follow the 30-nm fiber assembly. The highest condensation reaches at the metaphase chromosome (see Sects. 2.4 and 2.7) with each chromatid having an estimated diameter of 700–750 nm (Fig. 2.4). A hierarchical folding of chromatin is believed to take place from the 30-nm through the 700–750 nm fibers by progressively looping the 30-nm fiber. In fact, different stages have been registered, such as the stages of fibers of 100–130 nm, 200–250 nm, and 500–750 nm [12]. The reverse process is expected from telophase to G1 leading to a progressive decondensation of the metaphase chromosome. During interphase, the chromosomes are heterogeneously compacted into distinct globular domains that follow the same principle of hierarchical chromatin folding into loops. However, condensed chromatin of transcriptionally inactive genes has been described in intermediate levels (100–200 nm). Loop folding depends on topoisomerase II and a set of proteins, collectively known as structure maintenance of chromosomes (SMC). The metaphase chromosome consists of arrays of loops linked to SMC scaffolds on an imaginary chromosome axis. Likewise, chromatin domains of transcriptionally inactive genes are due to SMC scaffold orienting high frequency of intra-chromatin looping. Along the interphase chromosome, 11-nm chromatin fibers (euchromatin) and fibers with increasing diameters ranging from 30-nm to 250-nm (heterochromatin) coexist.

2.4 Chromosomes

As seen, the ensemble of double-helix DNA molecules is invariably much longer than the human organism itself. The DNA's capacity of folding in a precise manner gives it the ability to fit in the cells' nuclei, further protects it throughout the cell cycle, and ensures a proper gene expression regulation at the level of transcription (see Chap. 4). The maximum condensation of the chromatin achieved during the

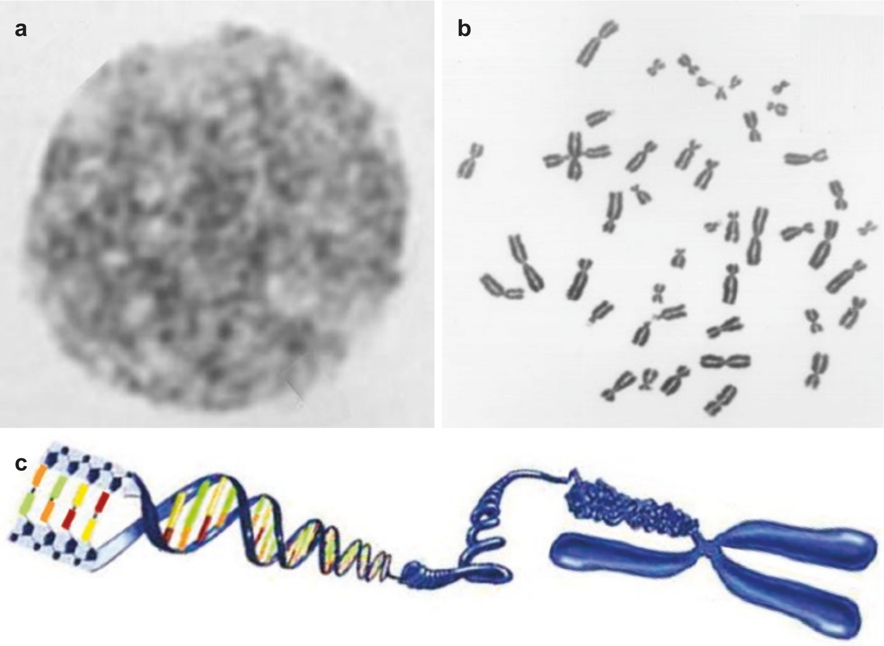


Fig. 2.4 Chromatin in an interphase cell showing homogeneous staining of the DNA inside the cell nucleus (**a**) and in a metaphase cell showing individualized chromosomes under a light microscope (**b**). These two stages of the genetic material reflect the under condensation in the interphase and the maximum condensation in the metaphase cell (**c**)

metaphase stage of cell division allows us to observe through an optic microscope several structures, called chromosomes.

The term chromosomes (from the Greek *chroma*, 'color' and *soma*, 'body') was originally coined by the German anatomist Heinrich von Waldeyer-Hartz, in 1888, to describe the basophilic filaments that could be observed inside the nuclei of cells. Some can currently employ the term in this wider sense, to refer to these portions of chromatin in the cells, formed by a single and continuous molecule of DNA, either if they are visible under the microscope or not. From now on, however, we will address it mostly to refer to those structures that can be microscopically visible as individualized pieces of chromatin due to their high condensation state during cell division (Fig. 2.4).

2.5 Number and Morphology of Human Chromosomes

The set of chromosomes that make up the genome is unique in regard to morphology, number, and the location of genes for each species. The study of chromosomes, their structure and inheritance officially started with Theodor Boveri and Walter Sutton, who independently stated, in 1903 [13], that genes lied at specific locations on the chromosomes, pointing out the thread-like structures as the cornerstone of genome organization. By combining cytology and genetics, Sutton also coined the term cytogenetics to refer to this new rising branch that was responsible for the study of chromosomes and their particularities.

Thereafter, the evolution of cytogenetics, and more specifically human cytogenetics, followed the improvements in microscopic lenses and chromosome preparation techniques that occurred in the following years. In 1956 [14], after much discussion on the topic, Tjio and Levan could determine that the correct number of human chromosomes in somatic cells was 46 (2n = 46). Since then, our knowledge has definitely evolved and, nowadays, cytogeneticists are able to study thoroughly the molecular biology of chromosomes, as well as the biological and pathological conditions related to them.

Each human nuclear chromosome is made up of DNA interacting with histone and non-histone proteins. Morphologically, the metaphase human chromosomes are ten thousand shorter than the stretched DNA and they are formed by the two copies of DNA resulting from the DNA replication in the cell cycle S-phase, the sister chromatids.

The sister chromatids are kept close together by a region with a highly repetitive DNA sequence, the centromere, which creates a primary constriction on the chromosomes (see Chap. 6). The centromere also plays a central role in the correct chromosome segregation during cell division, as it helps the connection between the chromosomes and the microtubules from the spindle. Besides that, it also divides the chromosomes into two different sections, called arms: the short arm (p—from the French *petit,* 'small') and the long arm (q).

Depending on the position of the centromere, each human chromosome can be morphologically classified: they are metacentric, when the centromere is located at the center, and divides the chromosome into two arms of similar length; submetacentric, when the centromere is near the center, dividing the chromosome into two slightly asymmetric arms; and acrocentric, when the primary constriction is positioned near one end of the chromosome, which produces a very short p-arm and a longer q-arm. Human autosomal acrocentric chromosomes (chromosomes 13 to 15, 21, and 22) also present a secondary constriction on their short arm. Due to this constriction, those chromosomes contain a segment that is visually distant from the rest of its body, the satellite, which is associated with the nucleolus formation and the production of ribosomal RNA (Fig. 2.5).

At the extremity of each arm, the chromosomes also present tandem repeats of a DNA sequence ($TTAGGG_{(n)}$), ranging from 3 to 20 kb, associated with specific proteins. This structure is called telomere and among other functions, it is responsible for the maintenance of chromosome structure and stability, protecting it from any terminal unwanted recombination (see Chap. 7).

There are also regions along the chromosomes that have an increased number of active genes with a high GC-content, also known as gene-rich regions, whereas other regions, as opposed to that, are gene-poor and enriched by AT-content. This is especially important because each chromosome is organized in its own pattern regarding these different regions, creating distinguishable and intercalating chromosome bands, after some specific staining procedures, which assist the chromosome identification and organization for the karyotype.

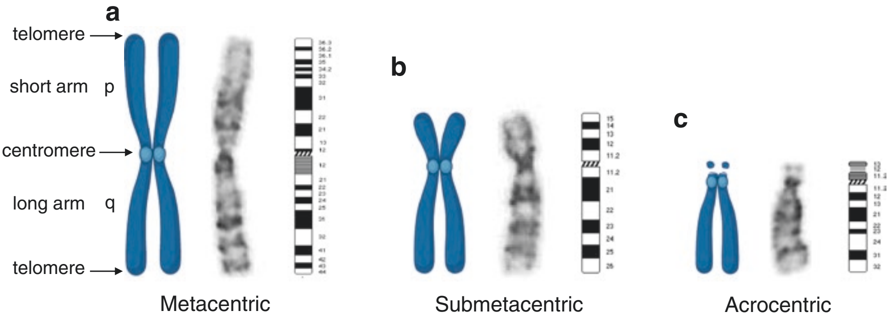


Fig. 2.5 Scheme of chromosomes showing the short (p) and long (q) arms limited by telomeres and centromeres. Examples of the metacentric chromosome 1 (**a**), the submetacentric 10 (**b**), and the acrocentric 14 (**c**) showing a simplified scheme, a chromosome with G-banding, and its ideogram, from left to right. The ideograms of the chromosomes show the G-bands numbered from the centromere in both arms

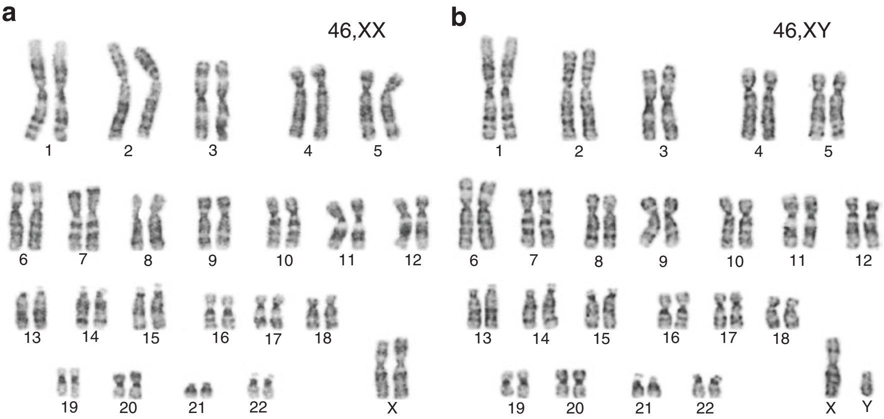


Fig. 2.6 Normal female (**a**, 46,XX) and male (**b**, 46,XY) G-banding karyotypes

2.6 The Normal Karyotype

The karyotype is the chromosome complement, or the whole set of chromosomes, of an individual. In humans, it is formed by 46 chromosomes, divided up into 23 pairs. Twenty-two of these pairs are identical in both males and females and are called autosomes, which are roughly numbered in a descending order of size, from 1 to 22. The other pair is constituted by two different chromosomes, termed X and Y, also known as sex chromosomes. This pair of chromosomes, unlike the autosomes, is different between males and females: females have two X chromosomes, whereas males have one of each sex chromosome. The female and male karyotypes are referred as 46,XX and 46,XY, respectively (Fig. 2.6).

The karyotype can be represented by ideograms, which are the chromosomes' schematic representation. The ideograms show the chromosomes' relative sizes and banding patterns. The bands and sub-bands are numbered, taking the centromere as a start point, and they are used to identify and refer to specific regions of the short and long arms of chromosomes. The tumor suppressor gene *TP53*, for example, is located on the short arm of chromosome 17 at the band 1, sub-band 3, and sub-sub-band 1, or at 17p13.1 (Fig. 2.7).

2.6.1 Autosomal Chromosomes

The members of the chromosome pairs, also known as homologous chromosomes, typically have the same subset of genes in the same order, one chromosome coming from the mother (maternal chromosome) and the other one from the father (paternal chromosome).

Morphologically, the homologous chromosomes usually cannot be distinguishable from each other. However, the sequence of the genes from the maternal and paternal chromosomes can vary slightly. These variant forms of each one of homologous genes are called alleles, and the combination of the maternal and paternal alleles represents the genotype**.** The autosomal alleles usually interact in distinct ways at a functional level, contributing to the expression of a somatic trait, or a phenotype, i.e., the visible characteristics and traits of an organism.

The human *ABO* gene, for instance, is a good example of how alleles interact to express a phenotype. This gene, located on the autosomal chromosome 9, is responsible for the human blood group determination, and has three possible alleles: I^A, I^B and *i*. Two of them (I^A and I^B) encode a different glycosyltransferase that modifies the H antigen on the red cells. While the I^A and I^B alleles can be both expressed at the same time, i.e., they are said to be codominant, the allele *i* results in no antigenic protein and is fully recessive to the other two alleles, meaning that the presence of either one of the two alleles surpasses the expression of the phenotype linked to the allele *i*. One person can only have two of the three alleles, whose interaction determines their blood type: A, B, AB, or O.

2.6.2 Sex Chromosomes

The last pair of human chromosomes is constituted by the sex chromosomes, which unlike the autosomes, are not numbered. Instead, they are known as X and Y chromosomes. In women (XX), the pair of sex chromosomes is constituted by two homologous chromosomes. In men (XY), on the other hand, the chromosomes of the last pair are non-homologous. The X and Y chromosomes differ from each other

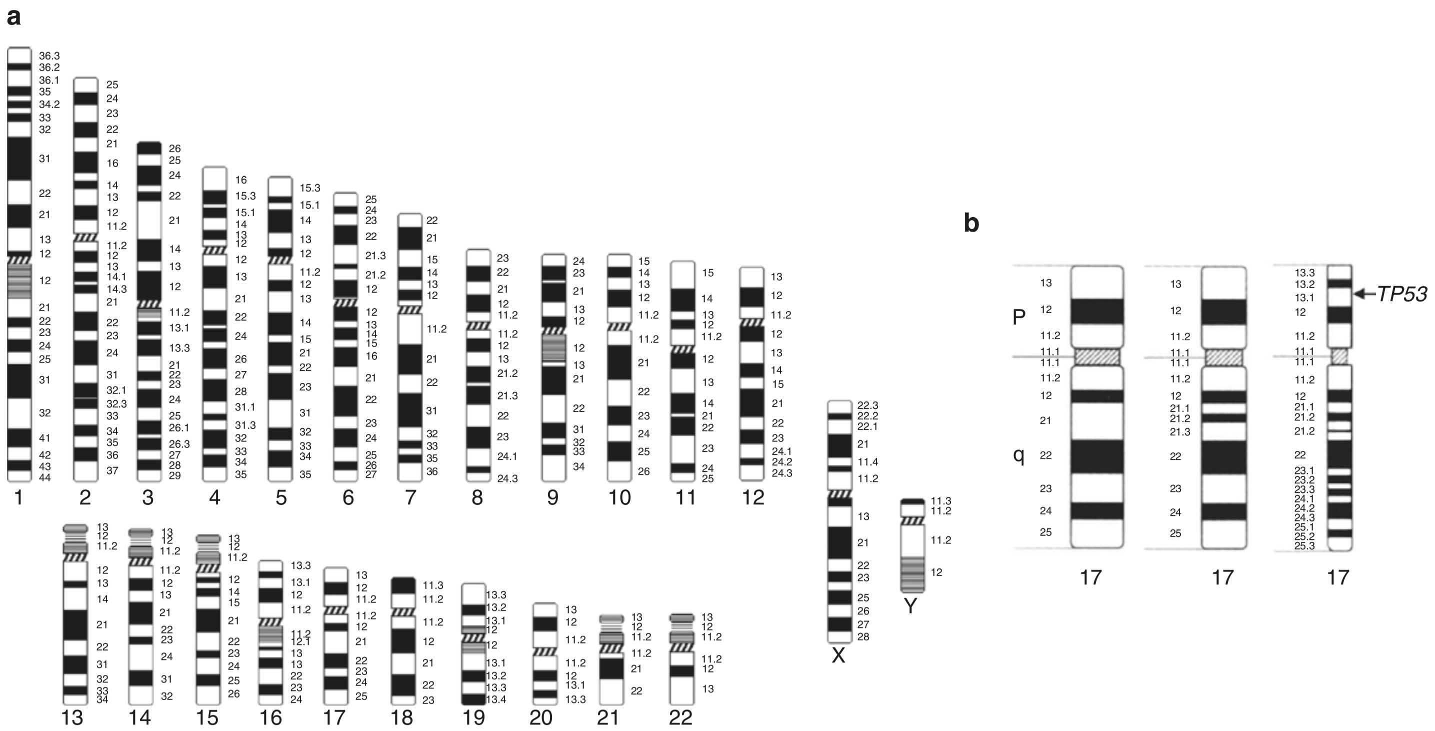


Fig. 2.7 Ideogram of individual chromosomes showing the G-banding pattern, with about 400 bands per haploid set (**a**); G-banding patterns of the chromosome 17 at different stages of condensation showing bands and sub-bands, and the position of the *TP53* gene at 17p13.3 (**b**)

in size and gene content, despite its common genetic background (both originated from an ancestral autosomal pair).

The Y chromosome is a small acrocentric chromosome and it is marked by a heterochromatic region on its long arm, with over 50% of its sequence being composed of repetitive elements. Regarding its gene content, it harbors only 63 genes, most of them being related to sexual development, including the *SRY* (Sex Determining Region Y) gene, responsible for triggering male development and regulating sex-linked traits. Besides Y-chromosome genes, autosomal genes can also play a role in sex-determination which is the case of the *SOX9* (SRY-Box Transcription Factor 9) gene, located on chromosome 17, whose function is crucial for male sex determination as well.

The X chromosome, on the other hand, retains many characteristics of autosomal chromosomes and it contains over 800 genes, which are not only related to sex development, but also control major somatic characteristics, such as neuronal development. Despite not being homologous, the X and Y chromosomes present regions of homology between them in their distal portion, also known as pseudoautosomal regions, which can pair and recombine during meiosis.

To balance this difference between women and men regarding the pair of sex chromosomes and its gene content, one specific mechanism, described by Mary Lyon, in 1961 [15], occurs early in the development of every somatic cell of females to achieve dosage compensation: the X-chromosome inactivation. This process is mediated by the X chromosome itself, more precisely by the X-chromosome inactivation center and its transcripts, and occurs early in the female development, ensuring that men and women have only one functional copy of the X chromosome in each of their somatic cells. Through this dosage compensation mechanism, most part of the genes from the maternal or paternal X chromosome is transcriptionally silenced through an enrichment of inactivation marks on it. Once silenced in one cell, the X chromosome maintains such state throughout this cell's clonal expansion.

The casual nature of the choice of which chromosome will be inactivated in women leads to some somatic cells expressing genes from the paternal X chromosome and other cells from the maternal one, characterizing women as true mosaics. The silenced X chromosome is usually more condensed, presenting a facultative heterochromatin, and can be distinguishable in the interphase nuclei, as a dark body, smaller than the nucleolus, called Barr body. The Barr body was initially described in female cats' neurons in 1949 and, later on, due to the establishment of Mary Lyon's theory, its presence within cells with more than one X chromosome was associated with the inactive X-chromosome.

2.7 Chromosomes and Cell Division

Cytogenetics also focuses on studying the behavior of chromosomes during cell division, either if it aims at growth, development or cell turnover in somatic cells, a process called mitosis, or at reproduction, in which germ cells undergo a process called meiosis. Mitosis and meiosis have distinguishable aims, but chromosomes are key elements for both. Since chromosomes keep DNA tightly packed, they help it remain uniformly distributed in the nucleus, intact and accurately segregated in cell division.

Mitosis is the process of cell division that ensures the maintenance of an organism by assisting its development and the cellular turnover within the tissues. It also guarantees the preservation of the genetic identity of somatic cells, as it results in two cells with the same genetic background and the same number of chromosomes. The word mitosis, coined by Walther Flemming, in 1882, derived from the Greek *mítos*, and means 'warp thread', alluding to the thread-like conformation of chromosomes at the onset of mitosis.

The first stage of mitosis, the prophase, follows the phase in which the synthesis of DNA occurs (Fig. 2.2). By the time the cell cycle reaches the mitosis phase, the DNA is already duplicated and each one of the 46 chromosomes, which once consisted of a single molecule of DNA, is now formed by two identical molecules (the sister chromatids) that are kept close together by the centromere. During the prophase, the chromatin starts to condense, gradually forming spirals. The nuclear membrane disappears and the mitotic spindle, which is constituted by microtubules, is formed and will be responsible for the correct separation of sister chromatids. The following stage, the metaphase, is marked by the highest condensation of the chromosomes, their link with the spindle through the centromere and their equatorial alignment in the cell. Due to its high condensation state, the chromosomes are better visible through optic microscopes at this stage. Therefore, drugs that aim at stopping the progression of mitosis and arrest the cell at this phase are commonly used in routine chromosome preparations.

The anaphase follows the metaphase and is marked by the separation of sister chromatids due to the shortening of spindle fibers. This leads to each one of the sister chromatids migrating to the extreme poles of the cell. After that, the chromosomes become more diffuse and the nuclear envelope is reconstituted, which marks the last mitosis stage, the telophase. The cytoplasmatic division, or cytokinesis, follows this stage and marks the end of mitosis, separating the original cell into two daughter cells with equal amounts of DNA (Fig. 2.8).

The meiosis (from the Greek *meiosis,* 'lessening'), on the other hand, occurs in sexually reproducing organisms and aims at generating gametes, cells with half the number of chromosomes and the amount of DNA (in humans, n = 23) presented by the precursor cell. The number of chromosomes is reduced to guarantee that zygotes, after fertilization, have the correct diploid number of chromosomes (2n = 46). Meiosis is the key process through which genetic information is passed on from an individual to their offspring and the events that take place during this process ensure the balance between the maintenance of the genetic information and the recombination of chromosomal portions from maternal and paternal chromosomes to warrant variability and the generation of a new individual.

The meiosis can be divided up into two major processes that follow the replication of DNA: meiosis I and meiosis II. Just like mitosis, both have four distinguishable stages termed prophase, metaphase, anaphase and telophase, with very similar

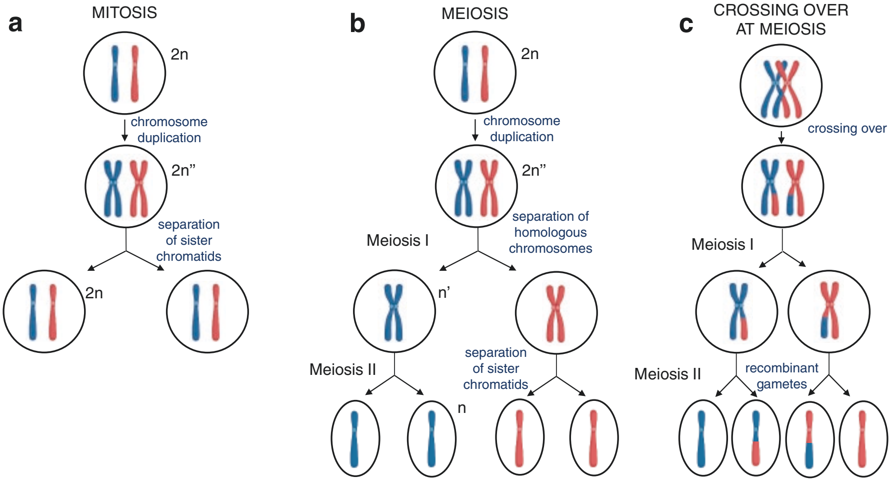


Fig. 2.8 A simplified scheme of the main steps of mitosis from a diploid cell (2n) producing two diploid cells like it (**a**), and of meiosis with two cell divisions producing haploid (n) gametes. Scheme of a crossing over between chromatids from the paternal and maternal chromosomes, showing some of the many possibilities of recombinant gametes (**c**)

events. Meiosis I is considered to be reductional because it segregates the two homologous chromosomes, generating two haploid cells (n = 23), with each chromosome containing two sister chromatids.

The longest phase of meiosis I is the prophase, which is divided up into five stages: leptotene, zygotene, pachytene, diplotene and diakinesis. The key event that occurs during prophase I, which is different from the mitotic prophase, is the alignment and pairing of homologous chromosomes (synapsis) followed by the homologous recombination, or crossing over, in which homologous chromosomes exchange genetic content. The crossing over is the source of genetic variation in the process of gamete generation. The exchange of genetic content between paternal and maternal homologous chromosomes produces gamete chromosomes that may differ from the chromosomes of the somatic cells in the same individual. From this point on, the homologous chromosomes with their pairs of sister chromatids attach to the microtubules of the meiotic spindle and are guided through different poles, and separated into two different cells. This event is followed by meiosis II, which is similar to the mitosis in somatic cells, but still, functionally different, as it involves the separation of the sister chromatids of the homologous chromosomes from the haploid cells generated in the previous phase. In humans, this results in four cells with 23 chromosomes each (Fig. 2.8).

Even though both processes are extremely controlled, errors in both the meiosis and mitosis can occur. Mitosis errors occurring in somatic cells, for example, can lead to cells with a different genetic content. Depending on the affected genes, these errors can lead to variable conditions, such as cancer. Meiosis errors, on the other hand, impact the offspring of the individuals. Errors in both processes can affect either singular genes or whole chromosomes, originating chromosome abnormalities, which are going to be explored in the next section.

2.8 Chromosome Abnormalities

Different chromosomal abnormalities, including numerical and structural alterations, can occur in humans. They constitute a frequent cause of miscarriage, birth defects, and intellectual disabilities. Chromosome alterations are found in 0.6 to 0.8% live born infants, in 25% of all miscarriages and stillbirths and, particularly in the first prenatal trimester, in more than 50% miscarriages. These values indicate a high occurrence of chromosome alterations in all conceptions, even though most of them are highly deleterious and not compatible with fetal survival. There are two main classes of chromosomal alterations: numerical and structural. While numerical chromosome abnormalities are due to alterations in the number of chromosomes in the cells, the structural chromosome abnormalities are due to chromosome breaks and joining of the breakpoints resulting in chromosome rearrangements.

2.8.1 Numerical Chromosome Abnormalities

Numerical chromosome abnormalities can be divided up into two groups: loss or gain of individual chromosomes (known as aneuploidy) and gain of whole sets of chromosomes (known as polyploidy). They are found in around 0.38% live births.

Aneuploidy results from errors in the number of chromosomes, when there is a loss or gain of one or a few entire chromosomes among the 46 chromosomes. Instead of the normal presence of two copies of chromosomes from a pair (disomy), the presence of only one copy results in a condition named monosomy, whereas three copies of chromosomes results in trisomy. While the only monosomy compatible with life is the monosomy of the X chromosome, rare cases with four (tetrasomy) or five copies (pentasomy) involving exclusively the sex chromosomes may occur. In fact, there can be errors involving chromosomes from every pair, but only a few aneuploidies are compatible with life. Only trisomy involving the autosomal chromosomes 13, 18, 21, and the sex chromosomes are potentially viable, but even these conceptions may be frequently lost during pregnancies.

Aneuploidy results from errors in chromosome distribution during meiotic or mitotic cell division. The nondisjunction of the homologous chromosomes in Meiosis I (both homologous chromosomes going to the same pole instead of segregating to opposite poles) or the nondisjunction of the sister chromatids in Meiosis II results in gametes that contain a higher or lower number of chromosomes. Alternatively, anaphase lag can result in chromosome loss in gametes. When these errors occur during meiosis, the fertilization produces an aneuploid zygote and the conceptuses will present abnormal chromosomes in all cells. The incidence of these meiotic errors and, consequently, autosomal aneuploidies increase with advanced maternal age. If these errors in chromatid disjunction occur post-zygotically, during a mitotic division, on the other hand, cells with different chromosome constitutions will be present simultaneously in an individual. This situation, in which two or more cell lines differ in karyotype in an individual or tissue, is named chromosome mosaicism.

In contrast, polyploidy is a type of numerical chromosome abnormality with the presence of extra haploid sets of chromosomes. Cells with 69 (3n) or 92 chromosomes (4n) are known as triploid and tetraploid, respectively. Triploidy with an extra set of maternal chromosomes can result from fertilization of a diploid ovum, due to errors in a meiotic division, by a normal spermatozoon, while triploidy with an extra set of paternal chromosomes can result from the fertilization of a normal egg by two spermatozoa (dispermy) or by a diploid spermatozoon. Tetraploidy can result from failure of normal cell division resulting in cells with four haploid sets of chromosomes. The great majority of triploid and tetraploid concept uses result in miscarriage, being triploidy one of the major causes of miscarriages. Rare cases are observed in live born infants, usually presenting mosaicism.

2.8.1.1 Numerical Chromosomopaties

Only some chromosomal alterations occur with an appreciable frequency in live born infants. Among the most frequently observed are trisomies for chromosomes 21, 18, and 13, monosomy X, and the trisomies for the sex chromosomes (XXY, XXX, XYY), whose main clinical and karyotype findings are presented. Their partial karyotypes are shown in Fig. 2.9. Polyploidy and molar pregnancies will also be presented. These numerical chromosome alterations can be detected by the karyotype exam.

Monosomy X

The only monosomy compatible with life is the monosomy of chromosome X, which results in Turner syndrome and occurs in 1 in 5000 live female births. Most patients (around 55%) with Turner syndrome presents karyotype 45,X in all cells, but monosomy X may be present in mosaic (10% cases) including 45,X, 46,XX and 47,XXX cell lines. Partial monosomy X can also be due to structural rearrangements such as isochromosomes (20%), deletions (5%), ring chromosomes (5%) and others (5%). Monosomy X has an estimated frequency of 1–2% of all clinically recognizable pregnancies, and only 1% of 45,X zygotes lead to live born infants. Interestingly, the clinical features found in live born and adult patients with Turner syndrome are not highly life-threatening. The clinical features of Turner syndrome are highly variable between patients, including signs and symptoms such as growth delay, short stature, webbed and short neck, short sternum, wide intermamillary distance, cubitus valgus and congenital heart defect. Primary amenorrhea, streak

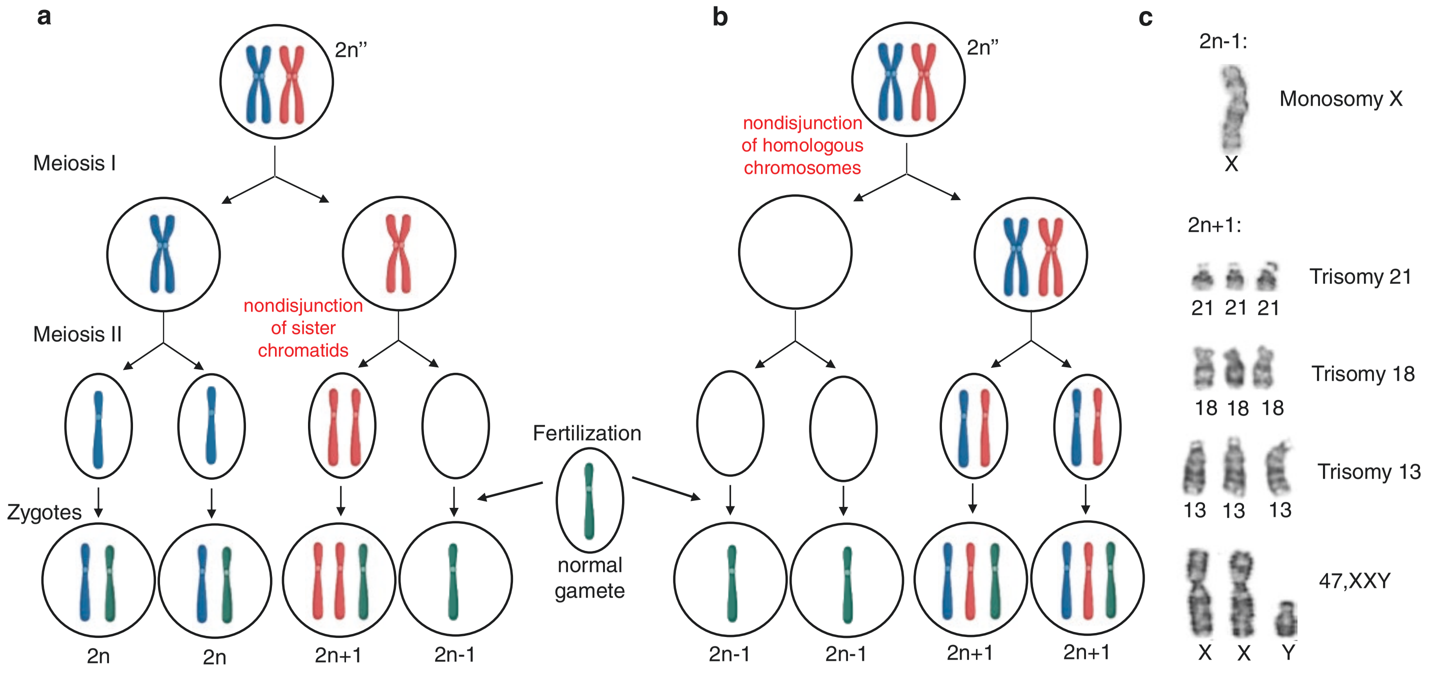


Fig. 2.9 Schemes of meiosis showing nondisjunction of sister chromatids at meiosis II (**a**) and nondisjunction of homologous chromosomes at meiosis I (**b**), from a diploid cell with their chromosomes already duplicated (2n”) and the zygotes originated after fertilization. Partial karyotypes from patients with monosomy X (Turner syndrome), trisomy of chromosome 21 (Down syndrome), 18 (Edwards syndrome), 13 (Patau syndrome) or sex chromosomes 47,XXY (Klinefelter syndrome) (**c**)

ovaries, delayed puberty and infertility are also found in females with Turner syndrome. Most affected girls and women usually present normal intelligence, although some of them may present learning disabilities.

Trisomy 21

Trisomy of chromosome 21 was the first known chromosome abnormality associated with a syndrome. Described in 1866 by John Langdon Down, it was named after him as Down syndrome. Trisomy 21 is the most frequent chromosomal abnormality in humans, affecting about 1 in 700–800 newborns. Most of the individuals with Down syndrome (around 95% cases) presents a chromosome constitution with an extra chromosome 21 in all cells, with karyotypes 47,XX or XY,+21. Mosaicism with karyotype 47,XX(or XY),+21/46,XX(or XY) are found in 1–2% cases. The remaining cases are due to structural abnormalities, especially Robertsonian translocations (see ahead). In pregnancies with trisomy 21 fetuses, increased nuchal translucency and absent or hypoplastic nasal bone are useful markers in the prenatal first-trimester ultrasound screening. The babies present muscular hypotonia, brachycephaly, and a characteristic facial appearance with a flat face, malar flattening, epicanthus, upslanted palpebral fissure, and protruding tongue. A short middle phalanx of the fifth finger, single transverse palmar crease, and short palm are also frequent findings. The patients usually present with short stature, congenital heart defects, and mild to moderate intellectual disability.

Trisomy 18

The incidence of trisomy 18 is about 1 in 7000 births, being more frequent in females, and resulting in Edwards syndrome. Most patients presents trisomy 18 in all cells, with karyotypes 47,XX or XY,+18, but mosaicism and partial trisomy due to structural abnormalities may also occur. The patients present intrauterine growth retardation, and low birth weight. Signs and symptoms include prominent occiput, a broad forehead, triangular face, hypertelorism, narrow mouth, microretrognathia, low-set, posteriorly rotated ears, and atypical position of fingers. Congenital heart defects, global developmental delay, and cognitive impairment are frequent features. Trisomy 18 is a life-threatening condition and fewer than 10% of patients survive for more than one year.

Trisomy 13

Trisomy 13 is found in about 1 in 12,000 births and results in Patau syndrome. The most frequent karyotype is 47,XX or XY,+13 but, in about 20% cases, it occurs due to trisomy in mosaic or unbalanced translocations. It is a severe disease with a limited life expectancy, with most patients surviving for only a few days or months. The patients present intrauterine growth retardation, muscular hypotonia, failure to thrive with severe global developmental delay, severe feeding difficulties, and apnea. Among the clinical findings are abnormalities of the fontanelles or cranial sutures, congenital heart defects, microphthalmia or anophthalmia, cleft lip or palate, low-set ears, and polydactyly.

Trisomy of Sex Chromosomes

The karyotype 47,XXY in all cells or in mosaic forms results in Klinefelter syndrome, which is present in about 1 in 1000 newborn males. The patients show tall stature, small testes, azoospermia or oligospermia with hyalinization and fibrosis of the seminiferous tubules, low serum testosterone, and elevated gonadotropin levels. Most is diagnosed in adulthood due to infertility. They may show gynecomastia, gynoid aspect of hips, sparse body hair, and usually present mild intellectual disability.

The 47,XXX syndrome, also called trisomy X or triple X syndrome, is found in 1 in 1000 live female births. Women with an extra X chromosome are usually taller than expected and without unusual physical features. They may present learning disabilities and delayed development of speech and language skills. Most females show normal sexual development and are fertile even though they may present premature ovarian insufficiency.

The 47,XYY karyotype has an estimated incidence of 1:1000 male births. There are no typical features of the double Y syndrome. Males, who are usually taller than expected, may show unremarkable signs and symptoms or may present learning disabilities, speech delay, and behavioral differences such as attention deficit hyperactivity disorder.

More rarely, tetrasomy or pentasomy involving sex chromosomes may occur with no known characteristic phenotypes. Alterations of the number of sex chromosomes cause less phenotypic effects and are more compatible with life. Considering the X-chromosome, only one of the X chromosomes must remain active in each cell, while the other X chromosomes are subjected to the inactivation of most of their genes and stay condensed during interphase. On the other hand, extra copies of the Y-chromosome may also have no effect since it is a poor-gene chromosome, as stated previously.

2.8.1.2 Poliploydy and Molar Pregnancies

Chromosome number alterations may involve not only one chromosome but also the whole haploid set of chromosomes. In triploidy, there is an extra set of chromosomes (3n) showing the karyotypes 69,XXX, 69,XXY, or 69,XYY. Triploidy has an estimated frequency of 1–2% of all clinically recognizable pregnancies, accounting for 15–20% of chromosomally abnormal first-trimester miscarriages. Almost all conceptions are lost during pregnancy with rare cases surviving to term, usually presenting diploid/triploid chromosome mosaicism. The outcome of triploidy depends on the origin of the extra set of chromosomes, indicating the occurrence of genomic imprinting (see Chap. 4). When the set is from the mother (2n maternal +1n paternal: digynic triploidy), the fetuses usually show severe intrauterine growth retardation and a variety of serious defects in the central nervous system, neural tube, heart, kidney, and limbs (syndactyly). Craniofacial abnormalities including relative macrocephaly, hypertelorism, and low-set malformed ears, cleft lip and palate, and micrognathia are also present. Usually, these pregnancies show oligohydramnios and a small, noncystic placenta. When the extra set of chromosomes is from the father (2n paternal +1n maternal: diandric triploidy), the triploidy usually results in molar pregnancy with partial hydatidiform mole due to abnormal trophoblastic tissue proliferation (both cytotrophoblast and syncytiotrophoblast) with vesicular swelling of the placental villi and some signs of embryonic/fetal tissues. In a complete hydatidiform mole, there is also over-proliferation of the chorionic villi, but no fetus material is present. The number of chromosomes is normal (2n) but both haploid sets of chromosomes have a paternal origin, indicating probable expulsion of the female pronucleus. These molar pregnancies are not viable and result in gestational trophoblastic disease with risk to metastasis. Polyploidy with four haploid sets of chromosomes (4n) is termed tetraploidy and show karyotypes 92,XXXX, or 92, XXYY. This infrequent chromosome alteration usually results in miscarriages with rare cases surviving to term.

2.8.2 Structural Chromosome Abnormalities

Structural chromosome alterations result from chromosome breaks followed by abnormal fusion of the breakpoints. In opposition to numerical chromosome alterations that are limited to few examples compatible with life, a great variety of structural rearrangements have been described. They are found in around 0.22% live births. This is expected since chromosome breaks can occur in different sites on different chromosomes. They can involve either only one chromosome, such as in deletions and inversions, or two or more chromosomes, such as in translocations. Thus, the fusion of the breakpoints may originate from a great diversity of structural rearrangements.

Structural chromosome rearrangements can be divided into two categories: balanced, and unbalanced. In balanced rearrangements, chromosome segments may be present in different positions but there is no loss or gain of genomic material. In unbalanced rearrangements, there is missing and/or additional genomic material.

Balanced structural chromosome rearrangements are found in about 1 in 600 individuals in the population and the carriers are usually phenotypically normal but are also at risk for miscarriages and/or children with unbalanced rearrangements. Since there is a loss and/or gain of genomic material in unbalanced rearrangements, phenotypical alterations are expected. The effect and embryo viability depend upon the size and genomic content of the unbalanced segment involved, being the gain of genomic material usually more well-tolerated than the loss.

The main types of chromosome rearrangements that can be identified by the karyotype exam are deletion (represented by the symbol del), duplication (dup), translocation (t), inversion (inv), and isochromosome (i).

2.8.2.1 Deletion

Deletions (del) result from chromosome breakages and the loss of a chromosome segment (Fig. 2.10). They can be terminal deletions (e.g. 46,XX,del(4)(p15)) or interstitial deletions (e.g. 46,XY,del(4)(p13p15)). In terminal deletions, neotelomere formation or telomere capture is needed for chromosome stabilization. Ring chromosomes (r), formed by breaks in both chromosome arms and joining of the broken extremities, result in deletion with loss of material from both arms. Deletions result in partial monosomies and the phenotype and viability depend on the size and nature of the deleted material. Deletions can occur sporadically due to meiotic errors or can be inherited from a parent carrying a balanced rearrangement. Thus, in cases of deletion, it is important to perform the parents´ karyotypes to establish the recurrence risk.

2.8.2.2 Duplication

Duplication (dup) is the gain of a chromosome segment and results in partial trisomy (Fig. 2.11). The extra material may involve a distal chromosome segment (e.g. 46,XX,dup(8)(pter p22)) or an interstitial segment (e.g. 46,XY,dup(1)(q22q25)). Duplications can be originated during gametogenesis, usually due to an unequal crossing over, occurring sporadically. They can also be inherited from parents with

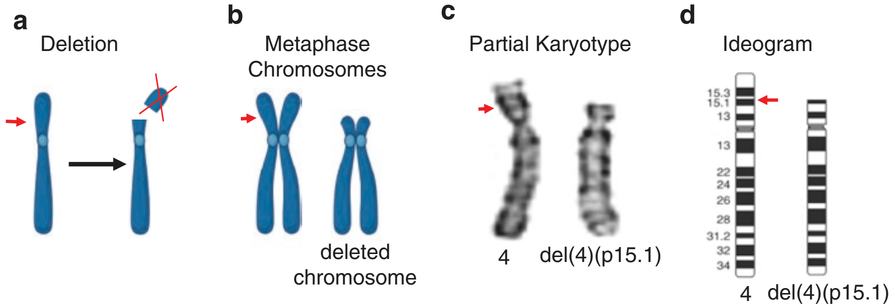


Fig. 2.10 Scheme of a terminal deletion with the production of a deleted chromosome after a break in the short arm (**a**), the pair of chromosomes with a deletion (**b**), and partial karyotype from a patient with karyotype 46,XY,del(4)(p15.1) (**c**), and the respective ideogram (**d**). Red arrows show the breakpoint position, in the normal chromosomes by convention

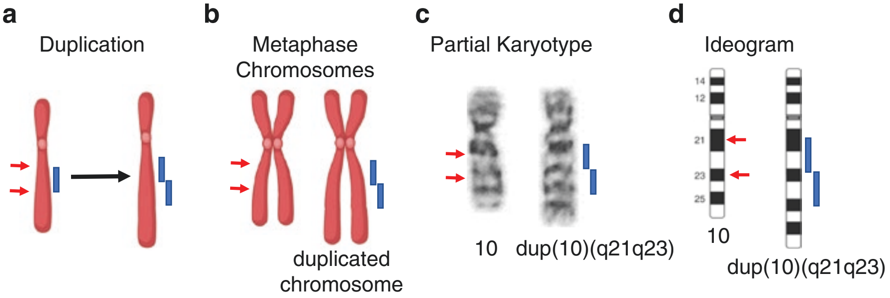


Fig. 2.11 Scheme of a duplication with the production of a duplicated chromosome (**a**), the pair of chromosomes with a duplication (**b**), partial karyotype from a patient with karyotype 46,XY,dup(10)(q21.3q23.31) (**c**), and the respective ideogram (**d**). Red arrows show the breakpoint positions, in the normal chromosomes by convention. The blue bars show the duplicated region

balanced rearrangements, and in these cases, a duplication is usually found together with a deletion in the same individual.

Extra material may also be found due to the presence of supernumerary marker chromosomes, which are small chromosomes with unidentified origin by karyotype exam. The most frequent marker chromosome is derived from chromosome 15 and is formed by an inverted duplication (inv dup(15)). Sometimes, the karyotype exam indicates the presence of an additional (add) material of unknown origin, whose identification will be provided by a parent's karyotype with a balanced alteration or by other exams, such as genomic arrays.

2.8.2.3 Translocation

There are two groups of translocations (t): reciprocal translocation and Robertsonian translocation, both originated from the exchange of segments between different chromosomes and, in the case of Robertsonian, acrocentric chromosomes. These translocations can be found as balanced rearrangements, usually with no phenotypic consequences, although the carriers are at risk for generating gametes with unbalanced rearrangements and, consequently, at risk for miscarriages and abnormal offspring.

2.8.2.4 Reciprocal Translocations

Reciprocal translocations (Fig. 2.12) represent one of the most common structural rearrangements, found in about 1 in 625 individuals in the population. They result from breaks in two different chromosomes with an exchange of fragments between them. These alterations of the position of chromosome segments result in a balanced rearrangement with no loss or gain of genetic material (e.g. 46,XX,t(1;11)

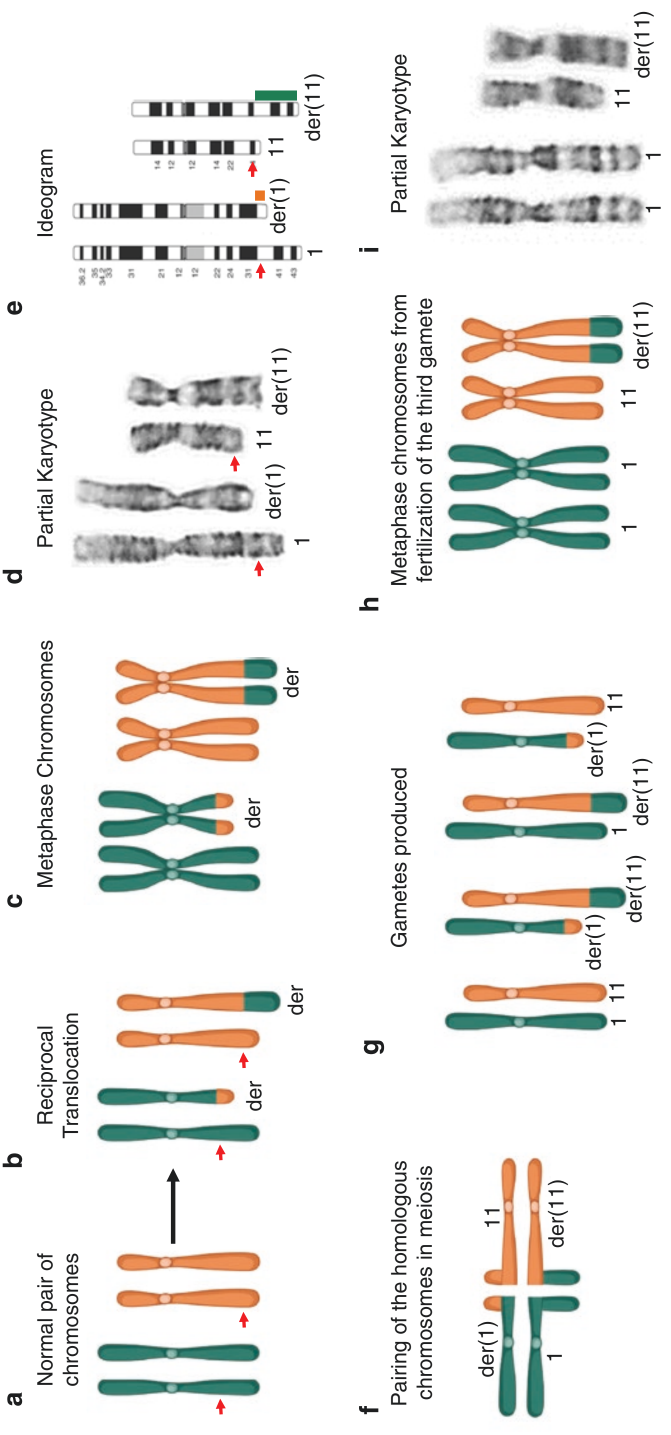


Fig. 2.12 Scheme of a reciprocal translocation with the production of two derivative chromosomes (der) with red arrows showing the breakpoint positions, in the normal chromosomes by convention (**a**, **b**); Metaphase chromosomes showing the reciprocal translocation (**c**); Partial karyotype from a patient with karyotype 46,XX,t(1;11)(q32;q25) (**d**), and the respective ideogram with orange and green bars showing the translocated regions (**e**); Pairing of the homologous chromosomes involved in a translocation t(1;11) forming a quadrivalent pachytene figure. During meiosis (**f**); Preferential gametes produced due to 2:2 segregation (normal, with balanced translocation, and two unbalanced options, in this order) (**g**); Metaphase chromosomes (**h**) and ideogram (**i**) of a patient with karyotype 46,XY,der(1)t(1;11)(q32;q25)mat, who inherited the derivative chromosome from a maternal balanced translocation resulting in 1q gain and 11q loss

(q32;q25)). Usually, there are no phenotypic effects for the individual. The translocated chromosomes are termed derivative chromosomes (der).

During meiosis, the normal homologous and the derivative chromosomes are paired and form a quadrivalent figure. During anaphase, different chromosome segregations may occur originating from several chromosome combinations. Normal and balanced gametes may be produced, with no phenotypic consequences being transmitted for generations. On the other hand, unbalanced gametes with loss and/or gain may also be produced (e.g. 46,XY, der(1)t(1;11)(q32;q25)pat), resulting in non-viable conceptions, miscarriages or alterations compatible with life but with clinical consequences (Fig. 2.12).

Usually individuals with unbalanced translocations present double chromosome imbalances, with the association of deletion in a chromosome involved in the translocation and duplication in the other. The risk for unbalanced translocations in the offspring depends on the segment sizes, gene content, and chromosomes involved in the translocation. The mean risk for abnormal offspring is estimated to be about 7% for female carriers and 3% for male carriers.

2.8.2.5 Robertsonian Translocations

Robertsonian translocations are found in about 1 in 800 individuals in the population and involve the fusion of whole arms of acrocentric chromosomes, after breaks on the centromere, or near the centromeric region. The carriers of a balanced Robertsonian translocation (Fig. 2.13) present 45 chromosomes with a derivative chromosome formed by the fusion of both long chromosome arms with the loss of their short arms (e.g. 45,XX,der(14;21)(q10;q10)). Since the acrocentric short arms contain only repetitive DNA sequences and redundant copies of ribosomal RNA genes, their loss has no clinical consequences. Balanced Robertsonian translocation carriers can be infertile, especially males, and are at increased risk for miscarriages

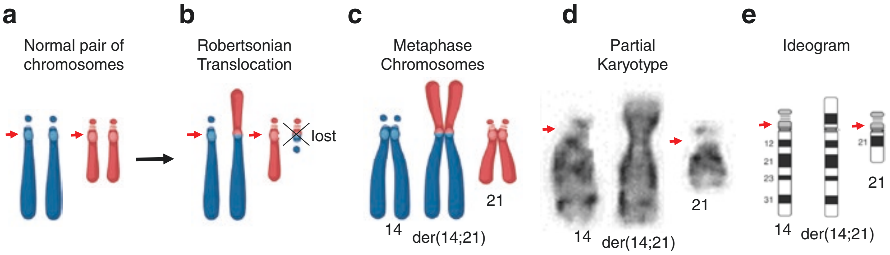


Fig. 2.13 Scheme of a Robertsonian translocation involving the acrocentric chromosomes 14 and 21 with the production of a derivative chromosomes (der)(14;21) and loss of the smaller segments (short arms) (**a**, **b**); Metaphase chromosomes showing the Robertsonian translocation with the fusion of the long arms of chromosomes 14 and 21 (**c**); Partial karyotype from a patient with karyotype 45,XX,der(14;21)(q10:q10) (**d**), and the respective ideogram (**e**). Red arrows show the breakpoint positions, in the normal chromosomes by convention

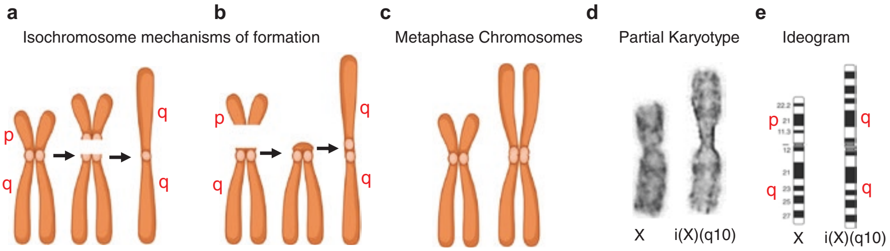


Fig. 2.14 Scheme of an isochromosome formation. Isochromosome formed by misdivision of the centromere with segregation of the two chromosome arms (**a**) or by breaks near the centromere with a U-type exchange and formation of an isodicentric chromosome (**b**); Metaphase chromosomes showing the normal chromosome and an isochromosome of the long arm (**c**); Partial karyotype from a patient with karyotype 46,X,i(X)(q10) (**d**), and the respective ideogram (**e**)

and chromosomally unbalanced offspring. In these individuals, after segregation during meiosis, from the trivalent structure formed by the chromosome pairing, different possibilities of gametes may be produced: normal gametes, gametes with the balanced translocation, and gametes with the unbalanced constitutions that result in conceptions with unbalanced karyotypes (e.g. 46,XY,der(14;21)(q10;q10)mat,+21). Among these, trisomy 21 due to Robertsonian translocations gametes represent about 4% of the patients with Down syndrome.

2.8.2.6 Isochromosomes

Isochromosomes (represented as i) are constituted by two copies of one chromosome arm (duplication), joined as a mirror image, and with no copy (deletion) of the other arm (Fig. 2.14). They may be produced by abnormal centromere separation during a cell division or by breakages in both chromosome arms, near the centromere, and joining of the breakpoints. A chromosome with two centromeres may also be produced and it is termed as isodicentric (idic). Isochromosome of the long arm of the X chromosome is a frequent finding (about 20%) in females with Turner syndrome (46,X,i(X)(q10)), but isochromosomes can also be found in rare cases involving autosome chromosomes (e.g. 46,XY,i(18)(p10)).

2.8.2.7 Inversions

Inversions are intrachromosomal rearrangements originated from two breaks in the same chromosome followed by an inverted fusion (180°) of the chromosome segment (Fig. 2.15). The inverted segment may involve the centromere (pericentric inversions), or just one chromosome arm (paracentric inversions). A pericentric inversion (e.g. 46,XY,inv.(8)(p12q23)) may result in a chromosome with a change

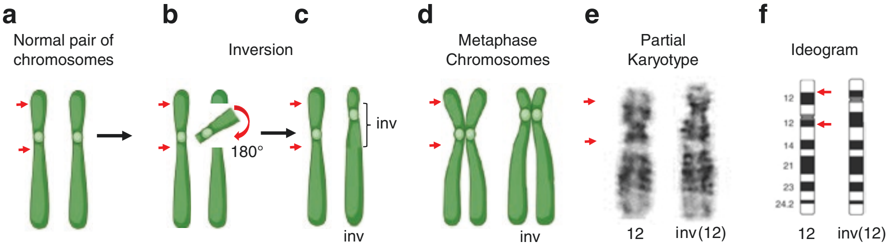


Fig. 2.15 Scheme of an inversion due to two breaks in each chromosome arm and fusion of the segment between the breaks after a 180° rotation (**a**–**c**); Metaphase chromosomes showing one normal chromosome and its inverted homologous (**d**); Partial karyotype from a patient with karyotype 46,XY,inv.(12)(p13q12) (**e**), and the respective ideogram (**f**). Red arrows show the breakpoint positions, in the normal chromosomes by convention

in the centromeric position. They constitute balanced rearrangements usually with no phenotypic consequences. However, even though this rearrangement involves only one chromosome, a peculiar pairing during meiosis, with the formation of a loop comprising the inverted segment for a correct homologous pairing, may occur. When a crossing over occurs at any site inside this loop, two recombinant chromosomes can be produced, each one with deletion of one of the segments on one side of the inversion and duplication of the other (e.g. 46,XY,rec(8)dup(8p)inv.(8)(p12q23)pat, and 46,XY,rec(8)dup(8q)inv.(8)(p12q23)pat). Thus, a carrier of inversions may produce normal gametes, gametes with the inverted chromosome, or gametes with the two types of recombinant chromosomes. The recombinant chromosomes are one with two copies of part of the short arm (dup p) and no copy of part of the long arm, and one with two copies of part of the long arm (dup p) and no copy of part of the short arm.

2.9 Some Considerations about Chromosome Abnormalities

The great majority of numerical chromosome alterations may be detected and identified by G-banding karyotype. Numerical alterations are easier to detect and may involve any chromosome. Monosomy for autosomal chromosomes, trisomy for most of the chromosomes, and triploidy are frequently lost during pregnancy. Only a few numerical chromosome alterations are compatible with life. Autosomal alterations result in a more severe phenotype than alterations in the number of sex chromosomes, and usually result in neuropsychomotor developmental delay, intellectual disability, dysmorphic features, and congenital malformations. Infertility is also a common feature in numerical alterations of the sex chromosome. Structural alterations, on the other hand, involve a great range of alterations considering the imbalance of different chromosome regions and sizes, and the phenotype is highly variable. Structural balanced alterations usually result in no phenotypic alterations unless the breakpoints interfere with gene function due to gene disruption or position effect, or due to undetected small genomic imbalances not identified by the karyotype. It is important to note that structural chromosome alterations may be sporadic (*de novo*) or inherited. Thus, in case of patients with unbalanced structural rearrangements (e.g. deletion, duplication, marker chromosome, derivative chromosome with unknown origin), it is important to evaluate the parents´ karyotype to determine if they present a balanced rearrangement, such as translocations or inversions. If so, they are at increased risk for offspring with unbalanced rearrangement, and genetic counseling is essential. Another group of patients in which a chromosomal evaluation is important is those individuals with atypical genitalia and disorders of sexual development since the karyotype exam can help the diagnosis of their condition.

2.10 Genomic Disorders

Genomic disorders constitute another group of genetic diseases characterized by genomic rearrangements in which the clinical phenotype results from the abnormal dosage of a gene(s) located within a rearranged genomic segment. The segments involved in the genomic rearrangements are smaller than 5–10 Mb but are larger than 10 kb. Microdeletions, microduplications, and some inversions are considered genomic rearrangements. Many of these are relatively frequent and recurrent since they involve unstable genomic regions. This high susceptibility to genomic rearrangements is due to the presence of low copy repeats (LCR), also known as segmental duplications, which are blocks of DNA longer than 1 kb in length with a high similarity (>90%). They are found in many copies throughout the genome but are usually clustered in certain genome regions. LCRs located <10 Mb apart from each other predispose to unequal crossing-over through a mechanism named nonallelic homologous recombination (NAHR), and results in changes of genome organization that can cause a loss or gain of genomic segments. Most of the genomic disorders are mediated by the presence of highly similar LCRs and are recurrent in the population, resulting in deletion, duplication, or inversion of genomic segments with similar sizes. One frequent recurrent genomic rearrangement involves a 3-Mb 22q11.2 deletion flanked by the LCR22-A and LCR22-D, which are LCRs found in the 22q11.2 region with a 98% DNA similarity, and results in the 22q11.2 deletion syndrome. Other genomic rearrangements that involve LCRs are the 1.6-Mb 7q11 deletion in Williams-Beuren syndrome, and the 4-Mb 15q11-q13 deletion in Prader-Willi syndrome. These genomic rearrangements are below the limit of microscope resolution and are not identified by G-banding karyotype, requiring other exams for their detection.

2.11 Methods for the Study of Chromosomes

2.11.1 Karyotype

The karyotype exam allows the identification of the human chromosomes and their abnormalities. It has been used for decades and is considered the gold-standard test to detect chromosome alterations ever since the identification of the first chromosome abnormality, the extra chromosome 21, in patients with Down syndrome in 1959.

The karyotype may be performed from different cells and tissues with nucleated cells that can undergo division since the chromosomes must be in their maximum condensation state, which occurs during cell division, in the metaphase, as mentioned previously. The most used material for karyotyping is lymphocytes from peripheral venous blood, collected aseptically in a tube containing an anticoagulant substance. To obtain chromosomes that are suitable for analysis, lymphocyte cultures are prepared adding a sample of the blood collected in a tube containing tissue culture medium supplemented with fetal bovine serum and phytohemagglutinin, which stimulates T lymphocytes to divide. The cells are cultured at 37 °C for about 72 hours. After this period, a mitotic spindle inhibitor, such as colchicine or colcemid, is added to the dividing cells in culture to arrest them in metaphase, allowing the study of the chromosomes. After the harvesting of the cells using a hypotonic saline solution and a fixative solution, the cells are spread onto microscopy slides. The chromosomes are stained, and the slides are analyzed using a light microscope. The staining method that result in G-banding chromosomes is the most used technique for chromosome identification. Under a light microscope, and usually with the aid of computerized image analyzers, the chromosomes are observed and organized to build the individual's karyotype and identify eventual chromosome abnormalities.

Additional banding methods can stain specific chromosome regions, such as the C-banding, which preferentially stains the constitutive heterochromatin of centromeres, the pericentromeric regions of chromosomes 1, 9, and 16, and the distal part of Yq; and nucleolar organizer region staining, which identifies these regions on the short arms of the autosomal acrocentric chromosomes.

Apart from peripheral blood lymphocytes, other cell types may be used for the karyotype exams, including skin fibroblasts, amniocytes, bone marrow cells, and tumor cells, among others.

Several chromosome alterations are not detectable by karyotyping when they are below the resolution level of G-banding karyotype, which is restricted to alterations involving more than 5–10 Mb. For example, microdeletions and microduplications, marker chromosomes, and complex chromosome rearrangements are not identified by karyotype analysis. Thus, other cytogenomic exams must be used, such as FISH and chromosome microarray.

2.11.2 FISH Test

Fluorescence in situ hybridization (FISH) is a method that combines classical cytogenetics with molecular tools. It is based on the hybridization of the DNA from the chromosomes, fixed on microscope slides, to specific probes (oligonucleotides) labeled with fluorochromes. The chosen probe for a specific exam is a DNA segment complementary to the region that will be investigated. The double-stranded DNA probe undergoes denaturation and is exposed to the patient's material on the slide, also denatured. When the DNA sequences of the probe and the target DNA are complementary, they anneal, and a hybridization signal can be visualized using a fluorescence microscope. The fluorescent signal, especially for detecting numerical chromosome alterations, can also be visualized in interphase chromosomes, with no need for cell culture. Thus, FISH can be used in metaphase and interphase chromosomes allowing its application in several samples, such as uncultured amniocytes and sections from paraffin-embedded samples.

A combination of FISH probes labeled with distinct fluorochromes can be used for different applications. For microdeletions, two probes are used: one for the target region and one for another region, as control. For the detection of the 22q11.2 deletion (Fig. 2.16), for instance, a specific probe (e.g. TUPLE1) for this region is used together with a control probe. In a normal individual, both chromosomes will show two fluorescent signals, while in the case of the 22q11.2 deletion syndrome, one of the chromosomes 22 will show two signals and the other just the signal for the control probe (e.g. 46,XY.ish del(22)(q11.2q11.2)(TUPLE1-)).

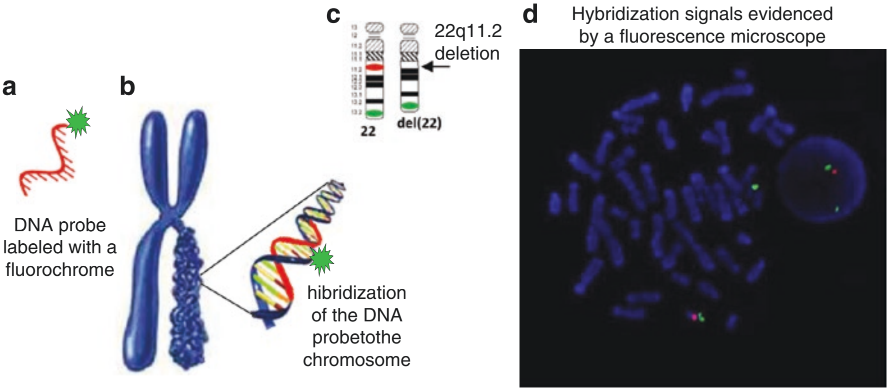


Fig. 2.16 Scheme of the FISH methodology. A DNA probe is labeled with a green fluorochrome (**a**), and it is hybridized to its complementary DNA on the chromosome, emitting a fluorescent signal (**b**). Two probes used for the 22q11.2 deletion identification: one for the target DNA (labeled in red) and one as control (labeled in green) (**c**). Metaphase cell and an interphase nucleus, each one showing two signals from the control probe and one signal from the probe specific for the 22q11.2 region, revealing the deletion (**d**)

The main limitations for this exam are the restricted number of probes used simultaneously in the exam and the need to have a suspected diagnostic hypothesis to choose a suitable probe. A pool with centromeric probes, labeled with different fluorochromes, is useful for prenatal and preimplantation diagnosis for the simultaneous detection of alterations of the chromosomes that are most frequently involved in numerical alterations (Chromosomes 21, 13, 18, X and Y).

2.11.3 Chromosome Microarray

Chromosome microarray, genomic array, comparative genomic hybridization array (CGH-array), and single nucleotide polymorphism array (SNP-array) are different nomenclatures for the test that allows the detection of losses (deletion) or gains (duplication or amplification) of genomic segments. It has been proposed to be the first genetic diagnostic test to be performed in patients with unexplained developmental delay/intellectual disability and/or multiple congenital anomalies since it offers a much higher diagnostic yield than the G-banding karyotype.

The exam is based on the hybridization of the test DNA to a slide or chip containing thousands, or millions of probes (oligonucleotides) organized in arrays in a specific order, corresponding to genome segments from the entire genome. Thus, copy number variations (CNV) from different parts of the whole genome can be detected simultaneously in one reaction, with high sensitivity (see Chap. 9).

One of the methods of chromosome microarray is known as CGH-array, since it is based on comparative genome hybridization (CGH), by the simultaneous hybridization of a test DNA and a control DNA (reference DNA), each one labeled with a distinct fluorochrome. The loss or gain of DNA segments is given by the over or underrepresentation of the specific fluorescent signals. When the array contains single-nucleotide polymorphism-based probes, with sequences corresponding to the two possible alleles of the SNPs, the test is known as SNP-array. The gain and loss of genomic material are given by the quantification of the hybridization signal considering the signals from appropriate internal controls, with no need for a reference DNA, unlike the CGH-array (Fig. 2.17). At present, most arrays contain SNP-based probes, in addition to CNV probes, since they may also provide reliable information about genomic rearrangement origin, uniparental disomy, and mosaicism, significantly improving the accuracy of the test. In all these methodologies, the slide or chip is read by a scanner, and the data analyzed using specific software to reveal genomic rearrangements. The test allows the identification of loss (e.g. 46,XY.arr 22q11.21(18648855_21800471) × 1, for a ~ 3-Mb 22q11.2 microdeletion) or gain of genetic material (e.g. 46,XY.arr[hg19]1q3 2.3q44(212508954_249224376) × 3, for a ~ 37 Mb duplication of the long arm of chromosome 1).

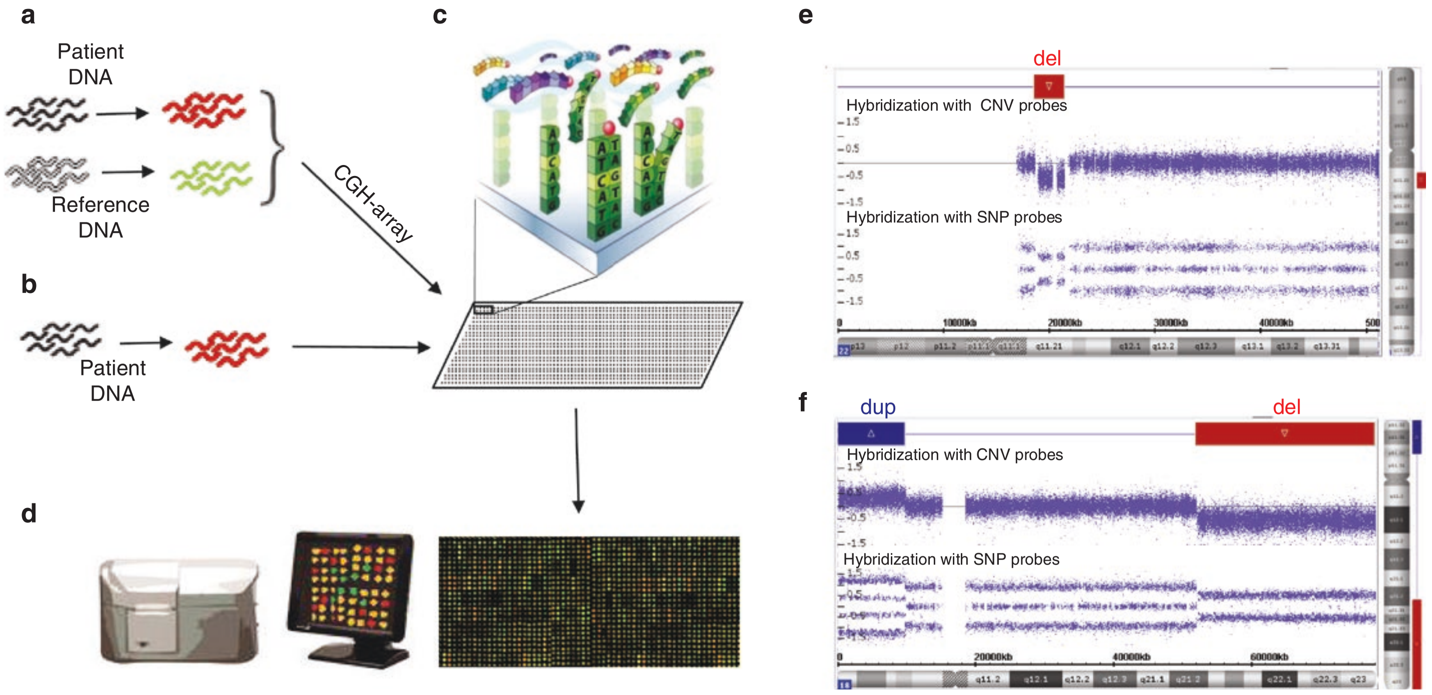

a
Patient DNA
Reference DNA
CGH-array
b
Patient DNA
c
d
e
del
Hybridization with CNV probes
Hybridization with SNP probes
10000kb
20000kb
30000kb
40000kb
f
dup
del
Hybridization with CNV probes
Hybridization with SNP probes
20000kb
40000kb
60000kb

Fig. 2.17 Chromosome microarray may be performed either using comparative genome hybridization with a test DNA and a reference DNA (**a**) or using only the test DNA (**b**) to hybridize to a chip containing probes for the whole genome (**c**). Using a scanner, the chip is read (**d**) and each signal is converted into organized information by a specific software that can detect loss (deletion) or gain (duplication) of genetic material (**e**–**f**)

Although the chromosome microarray allows the identification of submicroscopic losses and gains in the genome with a great resolution, there are some limitations. Balanced rearrangements, in which there is no copy number alteration, cannot be detected by this test. Also, the position of additional material is not provided by chromosome microarrays. The copy number variation is a frequent finding in the general population, being the most important class of variation in the human genome with many DNA segments differing in dosage between individuals. Therefore, the clinical interpretation of CNVs found in chromosome microarrays is not an easy task since there are several CNVs across the genome with no apparent association with disease phenotypes, being considered benign CNVs, and others, rarer, that are associated with phenotypic alterations, considered pathogenic. The interpretation of their pathogenicity is still a challenge for geneticists since the size, gene content, overlap with known benign, and pathogenic variants must be considered for the correct interpretation of the clinical impact of genomic variations.

2.11.4 Whole-Genome Sequencing

Numerical and structural cytogenomic abnormalities may also be identified by whole-genome sequence analysis. In large-scale DNA sequencing, DNA is cut in short fragments that are amplified and sequenced (see Chap. 3). Apart from allowing whole-genome DNA sequencing, by different methodologies, the over or under-representation of sequence reads of DNA segments can indicate losses or gains of DNA, also providing information referring to CNVs. Thus, not only numerical abnormalities but also structural abnormalities, such as deletions and duplications, can be detected. Balanced rearrangements can also be detected, by analyzing chimeric reads. The increasing efficiency and decreasing cost of whole-genome sequencing will probably allow it to be used in the future for clinical diagnosis considering the possibility of the simultaneous detection of mutations and copy number variations.

2.12 Cytogenomic Nomenclature

The reports of diagnostic tests must follow the standardized international nomenclature. A guide named ISCN: An International System for Human Cytogenomic Nomenclature [16], regularly updated since 1960, provides standardized criteria for the cytogenomic and molecular nomenclature for karyotype, FISH, microarrays, region-specific assays, and sequence-based assays. When using DNA coordinates, it is important to provide the genome reference assemblies by the Genome Reference Consortium used, such as the GRCh37 (also known as hg19), released in 2009, and the GRCh38 (also known as hg38), the assembly of the human genome released in December of 2013.

2.13 Final Remarks

The study of human chromosomes provides interesting insight into the genetic material, concerning DNA replication, chromatin structure and cell division. The karyotype exam with both autosomal and sex chromosomes organized in a specific order has an important clinical application, allowing the identification of numerical and structural abnormalities. Other exams such as FISH and chromosomal microarray can also be used for the detection of alteration in the genetic material. Since cytogenomic abnormalities can result in individuals with phenotypic alterations or miscarriage but also in normal phenotype, the use of different methodologies is essential for diagnosis and genetic counseling.

References

1. Watson JD, Crick FH. Molecular structure of nucleic acids: a structure for deoxyribonucleic acid. Nature. 1953;171:737–8.
2. Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia. 1950;6:201–20.
3. Wilkins MHF, Stokes AR, Wilson HR. Molecular structure of deoxypentose nucleic acids. Nature. 1953;171:738–40.
4. Franklin RE, Gosling RG. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. Nature. 1953a;172:156–7.
5. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature. 1953b;171:740–1.
6. Meselson M, Stahl FW. The replication of DNA in *Escherichia coli*. Proc Natl Acad Sci U S A. 1958;44:671–82.
7. Okazaki R, Okazaki T, Sakabe K, et al. Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. Proc Natl Acad Sci U S A. 1968;59:598–605.
8. Kornberg RD, Thomas JO. Chromatin structure: oligomers of the histones. Science. 1974;184:865–8.
9. Olins AL, Olins DE. Spheroid chromatin units (v bodies). Science. 1974;183:330–2.
10. Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science. 1974;184:868–71.
11. Finch JT, Klug A. Solenoidal model for superstructure in chromatin. Proc Natl Acad Sci U S A. 1976;73:1897–901.
12. Kireeva N, Lakonishok M, Kireev I, et al. Visualization of early chromosome condensation: a hierarchical folding, axial glue model of chromosome structure. J Cell Biol. 2004;166:775–85.
13. Sutton WS. The chromosomes in heredity. Biol Bull. 1903;4:231–51.
14. Tjio JH, Levan A. The chromosome number of man. Hereditas. 1956;42:1–6.
15. Lyon MF. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). Nature. 1961;190:372–3.
16. McGowan-Jordan J, Simons A, Schmid M, editors. ISCN 2016: an international system for human Cytogenomic nomenclature. Basel: S Karger; 2016.

Chapter 3
Methods to Study Genomic DNA Sequence Variation

Michel Satya Naslavsky and Marília de Oliveira Scliar

3.1 Introduction

Human genome variation is highly heterogeneous in scale, distribution across populations, and manifestation (from the molecular level to phenotype). This section will explore current methods that address such heterogeneity, their application regarding the objective of analyses, their advantages and limitations, and, finally, an overview of what is likely to come next. Although a description of historical facts is not the scope of this chapter, a brief reminder of the early developments illustrates the very fast pace rushed by genomic analyses from observations to direct experiments that take place in research and reach medical applications.

The broad variability in scale was empirically stated since the dawn of cytogenetic analyses combined with heredity studies. The rationale behind such a proposal is that variation in the chromosomal scale, observed in microscopy procedures, is relatively rare, and its occurrence is often associated with many clinical conditions (See Chap. 2). Therefore, the heritable phenotypic variability across individuals without major pathologies shall not be explained exclusively by large chromosomal abnormalities but rather from more subtle changes not detected by cytogenetic

M. S. Naslavsky (✉)
Department of Genetics and Evolutionary Biology, Biosciences Institute, University of São Paulo, São Paulo, SP, Brazil

Human Genome and Stem Cell Research Center, Biosciences Institute, University of São Paulo, São Paulo, SP, Brazil
e-mail: mnaslavsky@usp.br

M. de Oliveira Scliar
Human Genome and Stem Cell Research Center, Biosciences Institute, University of São Paulo, São Paulo, SP, Brazil
e-mail: mariliascliar@ib.usp.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_3

methods. By the end of the first half of the twentieth century, the determination of species-specific chromosome numbers and the systematic development of analytical methods to study chromosomes, such as karyotyping, naturally led to the comparison of a reference set with samples from patients. Lejeune, in 1959, proposed the correlation of the most common aneuploidy, the trisomy of chromosome 21, with the clinical features typical of Down syndrome [1]. Curiously, a few years earlier, the landmark paper by Watson and Crick describing the DNA structure was published giving rise to modern molecular genetic studies [2]. Across the 1970s, the development and expansion of indirect methods of measuring genetic polymorphisms through electrophoretic patterns in enzymes expressed in blood cells improved our comprehension of the distribution of variation in different populations [3]. Also, during this decade, the use of restriction enzymes, nucleic acid probes, and hybridization had an enormous impact in both detecting variability and pinpointing the genomic context of regions of interest, directly paving steps to genome-wide mapping. Towards the end of the decade, DNA sequencing by chain termination developed by Sanger and colleagues would begin a novel chapter in the sensitive detection of genetic variability [4].

The following decade saw profound advances in molecular biology. In 1986, Kary Mullis developed a method for DNA amplification *in vitro*, combining a pair of oligonucleotides (of which synthesis had been resolved just a couple of years before), dNTPs, DNA polymerase, and buffer to a series of temperature changes to optimize each step of what became ubiquitously known as PCR (Polymerase-chain Reaction) [5]. This method allowed precise amplification of specific genomic segments of interest and became an essential tool in most molecular biology protocols. Almost simultaneously, in 1984, the Alta Summit would be the incidental embryo of the largest initiative in human genetics: the Human Genome Project. The ambitious and expensive task would create a definitive reference map from which, to some extent, all projects could rely on and compare their results against [6]. As the project was approved but yet struggled to get funded and to convince the scientific community and society, until the final draft, delivered in 2001, many other advances were published, including the deposition of single-nucleotide polymorphisms and special program reports. One example is the 1993 report, which presented projects to develop and improve mapping, cloning, and assembly protocols, along with computational approaches and ethical implications. Therefore, the Human Genome Project had a pivotal role not only in delivering a reference human genome sequence but also in leveraging an entire ecosystem of research in genomics. Naturally, the next challenge to be tackled would be describing the enormous variant diversity discovered and their role in traits, including disorders with complete or partial genetic etiology.

Even though some current methods promise to approach the full spectrum of variation, it is still challenging to interrogate human genome variation using a single method. This broad range of variant categories also creates a practical problem of how to represent the human genome as a single reference, to which most detected alterations can be compared to, especially when considering population-specific (mostly rare) variation (See Chap. 11). In addition, it implies that for both research and clinical applications, there are several choices to be made which may, in turn, limit the observations and gradually bias the accumulated knowledge on a few classes of variation at the expense of others, due to cost, availability of analytical tools, and interpretation capacities [7].

3.2 Variant Categories

As mentioned before, the heterogeneity of variant types imposes limitations on each technique. Therefore, before choosing an analytical method, it is essential to understand what to expect (and not to expect) of variant categories to be interrogated in each technique. We can classify variants by at least three criteria: size, consequence, and frequency.

Among such criteria, the variant's length spectrum is key when choosing between two main groups of methods: fragment-based or sequence-based. While at the dawn of genetic analyses, it was a commonplace among scientists to think of variation as major chromosomal rearrangements, currently, due to sequencing techniques, the first type of variant that comes to mind is single nucleotide substitutions or short-ranged insertions and deletions (*indels*). Although both ends of the spectrum are true and relevant, the amount of variants carried by a population or an individual is likely to be asymmetrically distributed across variant sizes. Very large (over five million base pair—*bp*) genomic imbalances (insertions and deletions, commonly referred as copy number variants, CNVs; See Chap. 9), translocations and inversions are much less common than single nucleotide substitutions or short-range *indels* (up to 50 bp). These large events (over 50 bp) are called structural variants (SV), and there is a substantial range of sizes among them, roughly observed in an inverse correlation with its frequency (it is more common to find shorter than longer SVs) and to its type (it is more common to find CNVs than the other types of SV). As presented in Table 3.1, it is expected that about 4–5 million single nucleotide substitutions can be found on average per diploid genome; about a fifth of that

Table 3.1 Estimation of genomic variants per length category

Category of variant (Length-based)	Length of variant (bp)	Counts per diploid genome (Order of magnitude)	Total size of genome affected (Mbp)	Variants within coding regions
Single-nucleotide substitutions	1	3–5 million (10^6)	4–5	12 thousand
Short-range *indels*	1–49	180–800 thousand (10^5)	3–5	250
Structural variants	>50	4–5 thousand[a] (10^3) or 20–30 thousand (10^4)[b]	10–15	3[c]

[a] Short-read sequencing estimations [8]
[b] Long-read sequencing estimations [9]
[c] Loss-of-function structural variants [10]

corresponds to short-range *indels*; 10–100 times less frequently are SVs and, finally inversions, translocations and aneuploidies are the least common types of variants. On the other hand, the total size of the genome affected by the different types of variants presents the inverse trend (Table 3.1) [8–10]. Other two types of variants are widespread in the human genome, microsatellites (also known as short tandem repeats, STR) and mobile element insertions (MEI). STR consists of stretches of DNA composed of units of 2–15 nucleotides repeated in tandem (See Chap. 6), and because of their highly polymorphic nature they are very useful for DNA profiling easily obtained with standard PCR protocols. Mobile elements are DNA sequences that can change their number of copies or change their location within a genome, eventually affecting genes. MEI constitute approximately 50% of the human genome (See Chap. 8).

Depending on the genomic context, variants can be categorized according to their predicted consequence, which should ideally be validated by subsequent molecular analyses. Each predicted consequence can also be associated with potential changes in the function of the gene products or, alternatively, their direct effect on DNA interaction with regulatory elements.

Annotation tools can cross a variant file and diverse annotation datasets (most based on matched “CPRAs”: chromosome, position, reference allele, and alternate allele) to pinpoint diverse information important to predict variant consequence. If a variant is located within intergenic regions, inferring its consequence can be more challenging since annotations are limited by prediction of sequence-based regulatory motifs or a set of assays that evaluate the evidence of transcription activity, chromatin state, methylation of CpG clusters [11] and, recently, the method of Hi-C sequencing was developed and improved allowing detection of structural interaction of distant regions called ‘TADs’ (Topologically associated domains) and LADs (Lamina-associated domains) [12–14]. Such assays hold a promising contribution to genome annotation. The Encode Project Consortium is an effort to systematically improve the understanding of DNA elements and, subsequently, the effect of variants that fall along such regions [15].

When variants fall in regions defined by genes, there are substantially more annotation resources that can be helpful in categorization and inference of predicted consequence. The annotation informs if the variant is noncoding, intronic, UTR regions, or coding. For the latter, annotation informs which amino acid is affected and, if it is synonymous, nonsynonymous, stop gain, stop loss or start loss. Annotation can also flag potential splice sites based on relative location to the exon-intron boundaries and state whether it promotes a frameshift or in-frame (if the variant is an *indel*).

It is challenging to predict the functional impact of such alterations in their products (proteins or RNA): any given variant may have a neutral consequence; promote a gain of function by either increasing the amount or activity of the product, also known in genetics as hypermorphic; create a novel function or property (neomorphic), which could interfere in the other allele (dominant negative effect) or be expressed in a different tissue or moment (ectopic or heterochronic expression, respectively); finally, a variant can promote a partial or complete loss of function of the original product (hypomorphic or amorphic mutations). Among the latter, it is possible to infer with better precision its consequence, since premature stop codons, loss of start codons, frameshift indels and splicing motifs can be automatically annotated with reasonably high confidence, in addition to the detection of large insertions and deletions that span coding regions of the genes. However, there are pitfalls in automated annotations of potential loss-of-function (pLOF) variants when a frameshift variant has a nearby *indel* that restores the frame, or if a premature stop codon is located at the last exon (likely to activate nonsense-mediated decay pathway with the affected transcript). Recent algorithms such as LOFTEE (Loss-Of-Function Transcript Effect Estimator) address these putative outcomes [16]. Either way, it is remarkable how such annotations performed on large datasets of variants and allelic frequencies can improve our understanding of a given gene's intolerance to loss of function by measuring the observed number of pLOFs as compared to expectations based on transcript length and relative position, as a function of mutational saturation in datasets of more than 100 thousand individuals. Such metrics, named pLI (acronym for 'probability of being loss-of-function intolerant') and LOEUF ('loss-of-function observed/expected upper bound fraction') after further development of the calculations, are useful resources to estimate haploinsufficiency and the potential impact of variants assuming the gene's intolerance to inactivation, measured by the observed depletion of pLOF variants, describing genes associated with dominantly inherited disorders caused by hypomorphic function of the gene, which is not always trivial to infer in non-familial (also named sporadic) cases with variants originated by *de novo* events [16–18].

Finally, variants can be categorized by frequency. It is often arbitrary to establish frequency cutoffs and it depends on the application context. In population genomic studies, it is generally accepted that variants above a frequency of 0.5% or 1% in any given population can be considered common. Keep in mind the absolute number of counted alleles: even though the proportion is the same, 1 alternative allele in 200 alleles (0.5%) wouldn't be confidently tagged as common, as opposed to 0.5% calculated with 100 alternative alleles in 20 thousand alleles (10 thousand individuals, assuming a diploid locus). In molecular diagnosis of monogenic disorders, an upper bound cutoff of 5% can be applied for a stand-alone benign pathogenicity classification [19] and even very low allelic counts in control populations can provide supporting evidence of reduced pathogenic effect in causing Mendelian disorders.

Therefore, very large sequencing-based datasets enabled detection of a wider frequency range, including a set of very rare yet shared variants with allelic frequency as low as 0.005% (result of counting 10 alternative alleles in diploid loci of 100 thousand individuals, or 200 thousand alleles), which are useful in functional inferences such as LOEUF calculations and refinements in pLOF intolerance investigations. In addition, on the extreme of the spectrum, sequencing followed by annotation with large datasets can provide a high number of ultra-rare variants found in a single individual in heterozygous state (termed as singletons).

As more underrepresented a population is across large public datasets, the larger proportion of singletons can be identified in every sequencing project, given the sampling to avoid small degree relatedness of subjects. As a consequence, as sequencing initiatives containing diverse populations get larger in sample sizes, it is likely that the amount of singletons will eventually reflect a private set of variants shared only in families or lineages, including those that are *de novo*, that is, present in one individual but not inherited from either parent. Likewise, somatic variants usually detected in sequencing experiments from paired tissues would either fall in the category of mutational hotspots or *de novo*, besides falling by coincidence on positions that were previously detected in germline experiments. Such frequency spectrum promotes different types of methods for genomic analyses: while common variants can be interrogated in genotyping platforms (containing a selected list of previously known variable loci to be evaluated), rare and ultra-rare variants would only be detected by sequencing methods (without *a priori* hypotheses on what to find).

3.3 Methods in Genomic Analyses

As explained in the previous section, there are genomic alterations of various lengths, a fact that challenges the investigation of the full panorama of variation using a single method. Overall, depending on the length of variants, one method will be optimal over others to detect and describe variation, with high sensitivity and specificity, avoiding false-positives and, particularly, false-negative results. Most current widespread sequencing-based methods begin with random fragmentation of the source DNA in relatively short stretches and all methods rely on sequence alignment. In these cases, detection of duplications and deletions is not trivial, especially in heterozygosity. An alternative strategy, which in fact was developed before sequencing, is to analyze larger fragments of DNA. These methods rely on hybridization or conditional amplification and usually handle longer variants and complex rearrangements better than sequencing-based. We will explore some of these fragment-based methods in the following section and sequencing-based methods right after that. A secondary partition of the methods refers to targeted approaches versus genome-wide approaches, as the former requires some a priori evidence for interrogation of a certain variant, variant list or group of genes, and the latter is exploratory. A decision tree was built to help visualize this rationale (Fig. 3.1). In Sect. 3.4, we discuss some of the current applications, including a workflow for molecular diagnosis, rare-variant association testing, and polygenic risk scores. In Sect. 3.5, we present some promising perspectives, such as, cell-free DNA, long-read sequencing, and omics integration.

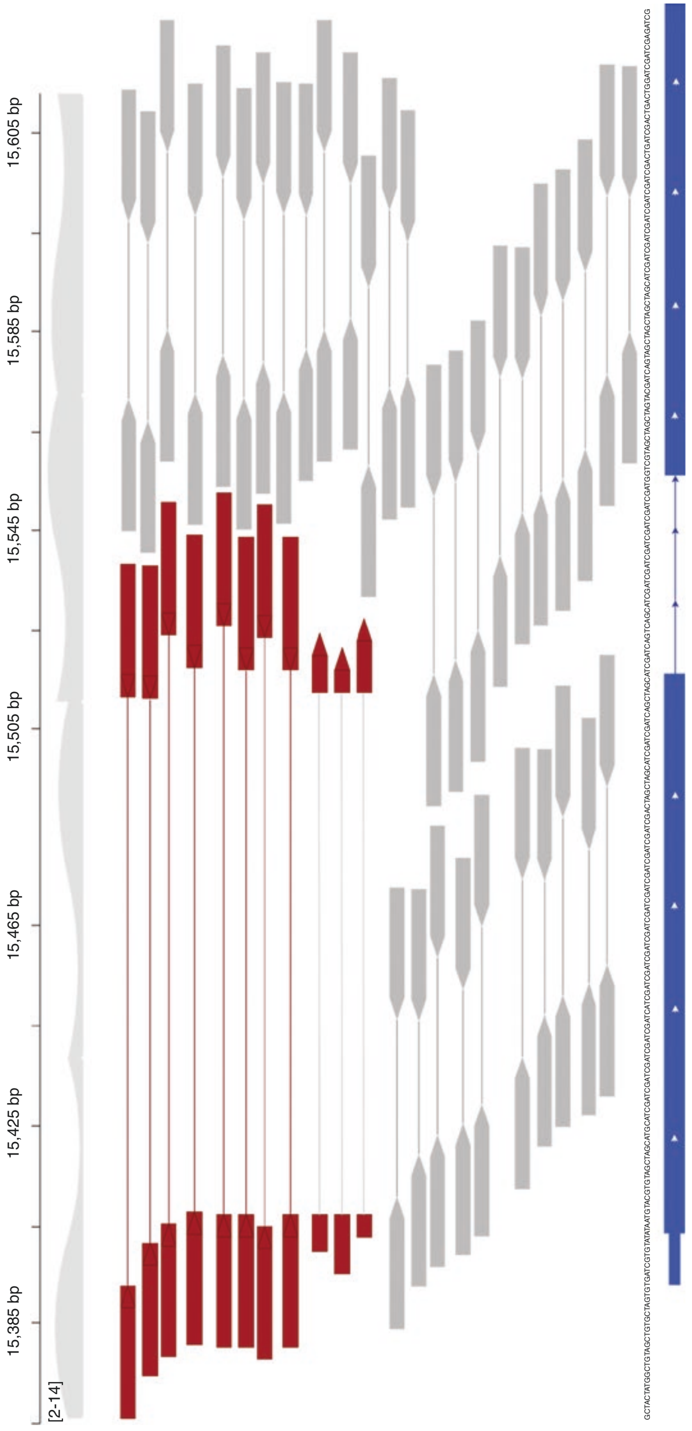


Fig. 3.4 BAM file with an example of paired-end clustering and split reads. One of the chromosomes had a deletion, which dropped local depth of coverage and promoted paired-end alignments to be more spaced than average. Also, a few reads were able to align to the reference after splitting. This example was adapted from a patient case from HUG-CELL (USP, São Paulo, Brazil)

3.3.1 Fragment-Based Methods

Large SVs are the cause of a significant amount of genetic disorders. In fact, there are more individuals affected by chromosome disorders than for all single-gene diseases. However, as mentioned before, there is a considerable amount of SVs, especially CNVs, per individual genome, indicating neutral or small effects of most variants. As variation grows in length, it becomes less common and more likely to be deleterious. Either way, screening the absence or presence of SVs, and quantifying them (in the case of multiple copies) is relevant in most genomic applications. In this section, we will briefly cover genome-wide and targeted fragment-based methods currently used in genomic analyses. Even though, by definition, whole-chromosome analyses fall within fragment-based methods, traditional karyotype and chromosome banding, observable under a microscope, will not be discussed here. We will cover fluorescence in-situ hybridization (FISH), array comparative genomic hybridization (array-CGH), multiplex ligation-dependent probe amplification (MLPA), and triplet repeat primed PCR (TP-PCR), that are vastly used methods in currently genomic analysis.

3.3.1.1 Fluorescence In-Situ Hybridization (FISH)

The introduction of FISH in the 1980 decade inaugurated the field of *molecular cytogenetics* that allowed locating specific DNA sequences on chromosomes and greatly expanded the sensitivity of chromosome analysis, becoming a powerful tool used in routine clinical diagnosis [20]. FISH experiment consists of using a fluorescence-labeled DNA or RNA probe capable of hybridizing to a complementary target sequence of a sample DNA. Probes can be labeled indirectly by modified nucleotides containing a hapten or directly by incorporating directly fluorophore-modified nucleotides. Further evaluation of signals under fluorescence microscopy reveals the chromosome location where the labeled probe binds, allowing the detection of various chromosomal abnormalities, including deletions, duplications, inversions, and translocations. The development of FISH came at the same epoch of the advent of the Human Genome Project that made available thousands of clone resources that could be used as probes [21]. One important advantage of FISH is its ability to perform analysis of interphase chromosomes, which allows the analysis of various samples, especially those from solid tumors that do not divide frequently (i.e., do not produce enough analyzable metaphases).

Since its development, many advances have increased the scope and sensitivity of the method. A powerful development was a 24-color karyotyping, called multiplex-FISH (M-FISH) and spectral karyotyping (SKY), in which each chromosome is painted with a different color, allowing a quick scan of all chromosomes to detect large deletions and/or duplications, translocations and complex rearrangements. However, site-specific probes are needed if more detailed information is required [22, 23].

3.3.1.2 Array Comparative Genomic Hybridization (Array-CGH or aCGH)

As explained above, FISH assays are suitable for investigating chromosome imbalances but rely on prior knowledge of which probes to use, one at a time. In contrast, the development of comparative genomic hybridization (CGH) allowed genome-wide screening for CNVs in a single experiment. CGH uses competitive hybridization between a patient and an unaffected control whole-genomic DNA (fluorescently labeled with different colors) to normal metaphase chromosomes. The fluorescence ratio of the patient and control hybridization signals along the chromosomes are then measured, revealing three possible outcomes: an equivalent signal, an overrepresentation or an underrepresentation of the patient's fluorescent signal [24]. Further development of the technique introduced array-CGH (aCGH), in which microarrays, consisting of a microscope slide with immobilized probes in defined positions, are used as targets instead of metaphase chromosomes [25]. The use of aCGH increased the resolution from 3 to 10 Mb of conventional CGH to 250 kb, and a higher density of probes can be used to increase resolution. Although the use of NGS-sequencing methods is increasingly replacing aCGH for CNV analysis in clinical testing and in research, at present, aCGH is the gold standard method to detect this type of variant and has been particularly useful in studying subtelomeric and pericentromeric rearrangements [26]. However, it is not appropriate for detecting other chromosomal abnormalities, such as inversions and translocation, that can be investigated by M-FISH or WGS.

3.3.1.3 Multiplex Ligation-Dependent Probe Amplification (MLPA)

MLPA is a rapid and cost-effective alternative to diagnose whole-exon CNVs on candidate genes [27]. The MLPA probe consists of two oligonucleotides, both containing the target sequence and a fluorescently labeled universal primer pair, identical for all probes. A stuffer sequence with a different size for each probe is attached to one of the oligonucleotides, giving each probe a unique length. Thus multiple probes can be hybridized simultaneously (multiplex). In the first step, the two oligonucleotides of each probe hybridize to immediately adjacent target DNA sequences. One oligonucleotide contains the binding site recognized by the forward primer; the other contains the binding site recognized by the reverse primer. Then, the pair of probe oligonucleotides that successfully hybridized are ligated, and only the ligated probes are amplified by PCR. Each fragment corresponds to a specific MLPA probe and generates a fluorescent peak that can be detected by capillary electrophoresis. By comparing the peak pattern of the tested sample with the pattern of reference samples, the relative change in copy number can be identified. MLPA can use up to 40 probes in a single reaction, in which each probe is generally used for each exon of a candidate gene. Thus, it is very useful for disorders, such as Duchenne muscular dystrophy, in which a substantial proportion of affected individuals have pathogenic deletions or duplications in a known gene.

3.3.1.4 Triplet Repeat Primed PCR (TP-PCR)

Trinucleotide repeats expansions are the cause of many genetic diseases, particularly neurological and neuromuscular ones (See Chap. 6). Standard PCR protocols are used to detect modest expansions. However, for large expansions (>100 repeats), an alternative method is necessary. Until the development of TP-PCR method in 1996, Southern blotting was the gold standard to analyze this type of variation. However, Southern blot is technically demanding, expensive, and has limited power to detect interrupted alleles, and then encouraged the development of TP-PCR [28]. TP-PCR uses an external primer flanking the repeat plus a primer that can randomly hybridize to multiple possible binding sites within the repeat, resulting in a ladder pattern on the fluorescence trace that enables the identification of expansions compared to samples used as a reference. The method allows identifying large expansions but cannot detect the exact number of repeats if this number is >50. TP-PCR was first developed to scan expanded alleles in myotonic dystrophy, but since then, the technique was validated for many other diseases, such as Friedreich ataxia (FRDA), Huntington's disease, and spinocerebellar ataxia type 3 (SCA3).

3.3.2 Sequence-Based Methods

By definition, methods that evaluate the presence and quantity of DNA fragments, and allow for quantification, irrespective of short-range variations in the sequence itself, were presented in the previous section. On the other hand, sequence-based methods are defined by the ability to interrogate or detect alterations across the sequence of particular DNA stretches. It means that even though fragmentation of DNA itself is often required as an initial processing step, or that the analyzed fragment will physically hybridize with probes, the main outcomes of these methods are the nucleic acid sequences themselves that allow detecting the variation on sequences when compared to a reference. In the following topics, we will cover Sanger sequencing, genotyping microarrays, and detail next-generation sequencing.

3.3.2.1 Sanger Sequencing

Although currently DNA sequencing far surpasses other biomolecules' sequencing in cost, ease, and, as a consequence, volume of generated data, in early 1970s, methods of protein and RNA sequencing were more advanced, although time-consuming. Nearly in parallel, Maxam and Gilbert's method of stepwise chemical cleavage of DNA molecule and Sanger and Coulson's method of DNA extension with chain-terminating nucleotides were successfully implemented in laboratories worldwide, both using fragments separation by electrophoresis [4]. The next development, known as shotgun sequencing, took place in the early 1980s and was extensively used in the Human Genome Project. This method targeted random clones of constructs containing libraries of samples of interest for sequencing and *a posteriori* computational reassembly of larger DNA fragments. Sanger's protocol would eventually prevail due to the improvement of the method with fluorescence-based automated machines in 1987.

Sanger sequencing is a reliable method of genomic analyses, targeted for regions of interest which are subjected to amplification or cloning. Therefore, even if a project is designed to cover a library of fragments generated by amplification or fragmentation followed by cloning, individual region of interest analyses will take place. For instance, all exons of a single gene are PCR-amplified or all fragments of a mitochondrial genome from a given tissue are cloned, physically paralleled Sanger reactions (one per plate well or tube) will be performed, generating individual electropherograms, which will be aligned to a reference sequenced or queried across a collection of sequences.

After amplification, products are purified to eliminate non-incorporated nucleotides and primers, and quantified for downstream steps. Sanger sequencing reaction consists of the extension of single strands (one primer is used) by incorporation of standard 2′-deoxyribonucleotides (dNTPs) complementary to the template strand and chain termination after incorporation of fluorescence-labeled 2′,3′-dideoxyribonucleotides (ddNTPs). Reaction parameters such as cycling temperatures, extension times and, especially, dNTP/ddNTP ratios are optimized to produce a library of DNA strands of different lengths with one nucleotide difference each. Each fragment from this library has a fluorescent dye brought by the 3′-end incorporated ddNTP. This reaction is then submitted to a high-density polymer matrix electrophoresis, usually in capillaries, to support the intended separation resolution of one nucleotide. Using a steady voltage, the process of differential migration of the fragments with optimal separation of fragments occurs towards the end of the capillary, where a detector is placed and converts fluorescence to bytes, including intensity parameters. The final result is one electropherogram per reaction, with roughly 800–1000 peaks that can be base-called for further analyses. One standard procedure is to cover the same region at least twice, in two different reactions, one for each strand (namely forward and reverse reactions). Depending on the project, a higher depth of coverage (also known as vertical coverage, meaning how many times a high-quality base is independently sequenced and called) is needed: the draft of the Human Genome Project, which was completed entirely with Sanger sequencing, was 5–10-fold [29].

3.3.2.2 Genotyping Microarrays

The use of hybridization techniques for analyzing nucleic acids started before sequencing technologies and basically consists of exploring the property of complementarity between base pairs of anti-parallel strands of DNA and RNA. As mentioned before for FISH and array-CGH, it is straightforward to observe the results of hybridization between a probe (the sequence we have prior information about) and the region we are interested in detecting/quantifying.

The ability to miniaturize the synthesis of oligonucleotide probes onto a solid phase (usually glass slides, in a process called photolithography), the implementation of improved digital cameras, and the growing knowledge on allelic diversity contributed to the development of ever-higher density genotyping microarrays, often called DNA chips [30]. The overall methodological workflow involves enzymatic fragmentation followed by end repair, adapter ligation, and PCR, enriching the sample in products of less than 1 kb. DNA probes in the chip harbor the selected SNPs in several positions (overlapping probes), and SNPs themselves are selected based on frequency and by location, usually between within two restriction enzyme sites 1 kb apart. Allelic detection by hybridization without this a priori step of size selection by amplification can produce non-specific calls that increase background noise. Although these steps are fairly similar between the two main commercial microarray platforms (Affymetrix and Illumina), each has their own specificities: Illumina uses a probe-linked beadchip embedded in the slides and has a single base extension; Affymetrix uses ligation, and several washes to remove less stringent hybridizations.

These tools were essential to decrease the costs of genome-wide analyses for several applications, from family-based segregation and linkage studies to large sample-sized genome-wide association studies (GWAS), since automation greatly increases the through-put and reduces experimental variability. Currently, commercially available microarrays include many options of high-density sets of variants (>500 k markers) and enrichment of clinically relevant variants or copy number variants; besides the possibility of some degree of customization. Outside basic research applications, most companies that offer direct-to-consumer testing for ancestry or disease risk alleles are microarray-based. It is important to consider that each microarray chip is designed to interrogate a list of polymorphic alleles previously detected by sequencing projects, which might be biased on their own. Some commercial microarrays were developed to include population-specific variants and to some extent contribute to studies on diverse populations. Many GWAS studies benefit from an increasing density of variants through imputation, in which unobserved genotypes are inferred by using haplotypes from reference panels [31].

3.3.2.3 Next-Generation Sequencing

Pinpointing the large frequency spectrum of genomic variants, from ultra-rare to common, is only achievable by directly sequencing the DNA. As mentioned above, Sanger sequencing method revolutionized genomic science by providing a reliable and reasonably automated protocol that could consistently deliver the nucleic acid sequence of stretches of 700–800 bp. The main limitation of Sanger is parallelization itself. The Human Genome Project public effort overcame this issue using the challenging, costly, and time-consuming solution of distributing the job among hundreds of facilities worldwide, while the private effort did a similar approach, except that the hundreds of machines were centralized in a single facility (improving optimization and reducing costs). Either way, sequencing a whole human genome by the end of the first published draft in 2001 was still priced in the order of magnitude of 100 million dollars.

The (Recent) History of Next-Generation Sequencing

During the late 1990s and early 2000s, emerging sequencing technologies evolved from the combination of microfluidics and molecular assays advances such as emulsion PCR, bridge PCR, and adapter ligation. Three main next-generation sequencing (NGS) platforms were released almost simultaneously in 2005 and 2006: 454 pyrosequencing (later acquired by Roche), Solexa sequencing by synthesis (SBS, later acquired by Illumina) and Agencourt sequencing by oligo ligation detection (SOLiD, later acquired by Applied Biosystems) [32].

Illumina prevailed as the more broadly used method, and its in-depth protocol will be discussed in this section. Before each method is briefly covered, a few important NGS parameters are presented. Considerations about them define the usefulness and cost-effectiveness of each method and quality standards to be observed during analyses. As mentioned previously, depth of coverage is the number of times a base is independently called (i.e., read counts overlapping a single base). Although there is no consensus on an optimum minimum depth, 10-20x is usually the aimed range, even though some applications like somatic mutation detection require deeper coverage. The second parameter is the horizontal coverage, meaning the genomic extension that the sequencing project is aimed at mapping. Both parameters can be planned during the experiment design. The third parameter is the read length, which is usually restricted by the sequencing method. Finally, also limited by the sequencing protocol, the sequence output measured in number of reads and megabases is a value expected by each protocol and sequencing machine. All parameters are useful when designing the experiment, including the ability to multiplex several samples per run, and to expect minimum values for quality control and downstream analyses.

Pyrosequencing protocol provided by 454 involved fragmentation of genomic DNA and ligation to adapters, which would be baited by beads, generating an immobilized library. These beads were then emulsified for optimal isolation (one bead per emulsion compartment), later distributed on a picotiter plate for sequencing cycles (adding polymerase and one dNTP at a time). Incorporated nucleotides would release pyrophosphates, which would cascade a reaction of ATP and luciferase-catalyzed luciferin oxidation, generating visible light. Each well would provide up to 700-bp reads (typically 500 bp), not far from Sanger sequencing and, therefore, minimizing alignment and assembly procedures. Sequencing output of the latest Roche 454 machines was about 14Gb. Homopolymer detection (contiguous nucleotides of a single base) is challenging in most NGS protocols and was particularly critical in 454 chemistry.

Applied Biosystems SOLiD methods also used beads with immobilized oligonucleotides complementary to adapters linked to DNA fragments to be sequenced. Fragments on beads were also amplified and spread onto glass slides in polonies (‘polymerase colonies’). The extension and detection, however, used a unique experimental design where 8-bp long oligonucleotides had four different fluorescent labels to four dinucleotides located at the 3′ end, while the remaining nucleotides of the probe were degenerate. When a dinucleotide is stabilized, a ligase catalyzes the phosphodiester bond, unextended strands are capped and the fluorophore removed with 3 bp cleaved at the 5′ end of the probe, allowing the next cycle. The dinucleotide color code can be decoded by repeating the cycles offsetting the initiation primer by 1 base pair (n-1, n-2, n-3, and n-4), tiling the probe-fragment complementarity and generating overlaps between each sequence of colors. Although a complex procedure, this dinucleotide color-coded overlaps allows each base to be covered twice per read, increasing accuracy without substantially increasing cost. Each SOLiD output per run as for the last available machine was 90Gb. Two major drawbacks for this method probably caused its resistance to use and later discontinuation: very short read length of 50–60 bp, promoting a reduction in alignment quality and computational complexity increase; and problems in sequencing palindromic regions, which form hairpins and reduces ligation of probes to a critical level.

Illumina’s Sequencing by Synthesis

Illumina has established the main adopted technology by overall markets using the sequence-by-synthesis (SBS) method developed and improved from a combination of the original patents by Solexa and Lynx technologies. SBS consists of fragmentation of DNA in regular-sized segments to be ligated with adapters. Usually, fragments are around 500 bp of length obtained from native DNA, but mate-pair libraries aim at 2–5 kb long inserts and linked-read assays start with high molecular weight DNA that is isolated prior to fragmentation (Chromium technology, 10X Genomics). In all cases, only the extremes are sequenced (paired-end method), but regular spacing (short and long-range) improves *de novo* assembly and the ability to detect SVs. Indexing the fragments during library prep consists of ligating oligonucleotides to the ends of the fragment, before adapters’ ligation. Indices play a barcoding role allowing libraries to be pooled together, therefore multiplexing the run with several samples. After sequencing, computational scripts for demultiplexing will reassign reads to individual sample identifiers. Both adapters have complementary surface-bound oligonucleotides in a structure called a flow cell. Fragments are denatured and each strand is annealed with the fixed oligonucleotides for extension, and complementary strands are synthesized, now covalently attached (with phosphodiester bonds) to the flow cell. Clusters are generated by bridge amplification, populating the surroundings of the initial fragment with copies. As sequencing cycles begin, these clusters will be detected and monitored by a camera to register the incorporation of each dNTP to the fragment. The key for each added dNTP is that they have a reversible property, blocking the incorporation of more than one dNTP per cycle and unblocking after washing for further extension (Fig. 3.2). Reads range from 50 to 300 bp, but for genomic purposes are usually set to 150 bp. The paired-end sequencing allows for another round of dNTP incorporation cycles (also 150 bp), extending the complementary strands and providing a total of 2×150 bp long separated by the remaining unsequenced spacer of the original sample. It is relevant to

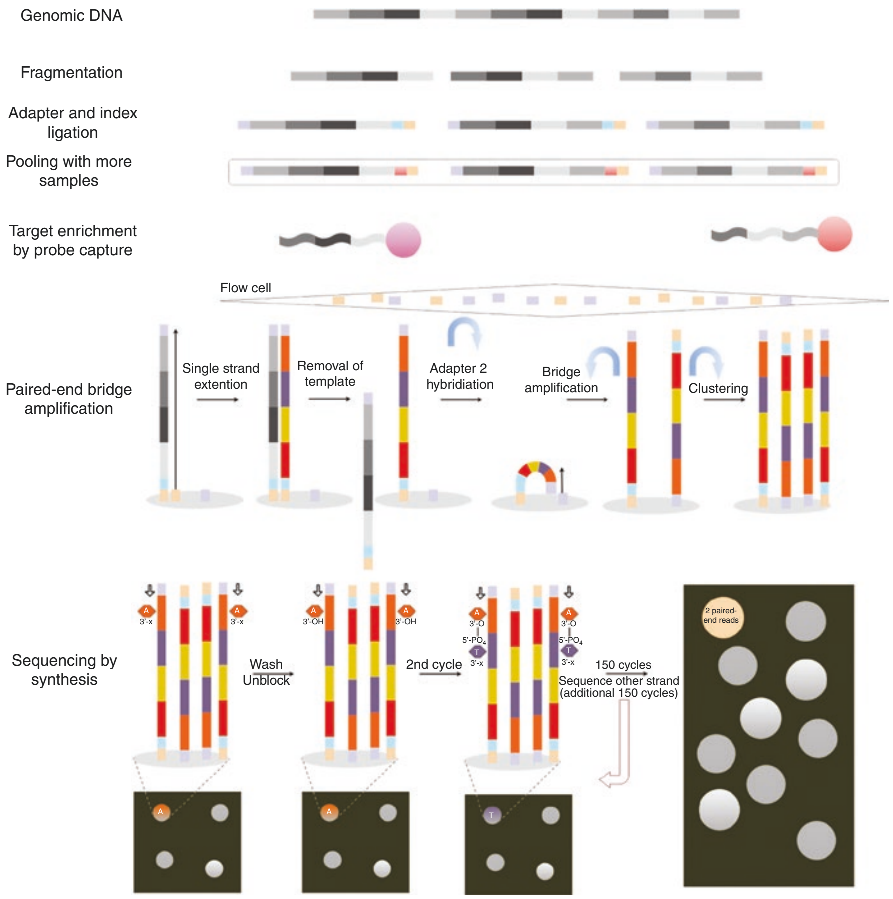


Fig. 3.2 Overview of NGS protocol (Illumina's sequencing by synthesis). Genomic DNA is fragmented in regular sizes (around 500 bp), which are ligated to adapters (lilac and light yellow) with indices (in this example, light blue for one individual and light red for another individual). Samples are pooled and optionally target enriched (in this example, carried out by probe capture attached to beads). Libraries are loaded onto flow cells containing oligonucleotide probes complementary to adapters. In any given probe, a fragment will be used as template for extension, and the original fragment is washed away. In a series of bridge amplifications, a cluster of fragments attached by both strands is generated. Sequencing by synthesis is performed in cycles, including incorporation of dNTPs that function as reversible terminators, and clusters are monitored for dNTP incorporation. Cycles restart to the reverse fragment within the same cluster, providing a paired-end read

stress that the paired reads are not reverse complements of each other, but rather the extremities of each fragment. The redundant representation of a region more than once (measured by depth of coverage) is, therefore, a result of independently sequenced reads from different fragments [32]. Output of Illumina SBS relies on a few parameters, some fixed and some adjustable: flow cell load capacity (there are different flow cell designs for each machine), cluster density (there is an optimal value), choice of single-end versus paired-end and sample multiplexing. Currently, Illumina offers a range of systems with outputs from 2 Gb to 6 Tb, which can be chosen in accordance with the project. In the clinical routine, it is a common practice to confirm findings using a different method, Sanger sequencing if the variant has short length. Sanger sequencing is commonly used to confirm or exclude the parent's carrier status.

Target Enrichment

As we have previously stated, the growing capacity of parallel sequencing millions to billions reads provides an opportunity to either go very deep on vertical coverage or very broad on horizontal coverage. In latest Illumina models (NovaSeq 6000), one can choose the higher output system of dual flow cells with 6 Tb output per run (and 20 billion reads). That means one human whole-genome (diploid genome of 6.4 Gb) could be sequenced at over 900x of coverage, or 900 whole-genomes could be sequenced at 1x. Although possible, both situations are not economically feasible, or actually desirable, for most applications. As we will see in the next section, whole-genome sequencing aimed at 30x is an accepted standard. But how would we make use of such a large output to make it both useful and cost-effective?

Besides multiplexing, which allows for pooling multiple samples along the same sequencing run, there are methods designed to enrich libraries with regions of interest that can be prioritized in sequencing experiments. Prior to the distribution of the libraries onto flow cells for cluster generation and sequencing, target enrichment methods can be performed. Two main strategies can be chosen to enrich libraries: probe-based capture or amplicon generation. In the first strategy, single-stranded DNA or RNA oligonucleotides are designed for the chosen regions of interest, synthesized, and attached to a solid phase surface such as a glass slide (resembling microarrays), or, more commonly, beads, followed by hybridization steps. Oligonucleotides probes must be long enough to allow for some mismatching, including indels, otherwise, allele dropout could be an issue (when one allele is not captured with equivalent success of the other allele). This also can be achieved by designing multiple overlapping probes spanning the region of interest (in a tiling setup), as some commercial options emphasize to be an improvement in enrichment by capture efficiency. The second strategy is to use multiplex amplification, generating amplicons of the regions of interest, using either primers or molecular inversion probes. There are advantages and disadvantages to amplicon-based enrichment. Customization and processing are easier and simpler than capture by hybridization of probes, and overall costs can be lower. However, there is a limitation on the number of amplicons that can be generated (amplicons that enrich whole-exomes were developed later and are usually more expensive than probes counterparts). Comparisons generally indicate that even with higher coverage, amplicon-based enrichment can be less uniform and provide a higher proportion of false-positives and false-negative results. However, for smaller panels and applications such as microbiome profiling, amplicon-based enrichment is widely considered [33].

Probably the most used application of NGS so far is target-enriched sequencing of human samples, specifically whole-exomes and gene panels. The combination of technological advances, reduction in equipment and reagent costs, multiplexing samples and target-enrichment led to an explosion of NGS data generation from the 2010s, both for academic purposes and clinical setups. Noteworthy is that the management and analysis of the incredible amount of data generated since then were only possible with the parallel development and advances of bioinformatics. Laboratories were able to standardize and streamline protocols to offer gene panels directed to groups of disorders such as hereditary cancer, neuromuscular and developmental disorders, costing no more than any complex health-related exam. Whole-exome sequencing (WES) had a pivotal role in the identification of genes associated with monogenic diseases, many of which required few family members to achieve probable candidates, a task that used to take time-consuming steps of STR profiling and linkage analyses. WES varies in terms of horizontal coverage, depending on how far into UTRs and introns the probes are designed to capture the target, but 120 Mb per diploid exome is a general value to consider. When aimed at 100x, the above-mentioned Illumina NovaSeq output could produce 500 WES per run, at a cost (reagents only) of about 100USD. Even adding other essential costs of equipment, computational resources, and, especially, high-skilled staff for producing and analyzing the data, WES will certainly stay as the gold standard for molecular diagnostics for a while [34, 35].

Whole-Genome Sequencing

Skipping the step of target enrichment and loading onto the flow cell the library of fragments ligated with adaptors and indices will produce the once holy grail of the scientific community worldwide: sequencing the whole human genome. The 1000 Genomes Project was launched in 2008 with the ambitious effort of sequencing thousands of individuals from 26 populations. Phase 1 included low-coverage whole-genome sequencing (WGS) of 179 individuals, WES of nearly 700 individuals and high-coverage WGS of two trios. The project was able to deposit a large number of variants previously unknown and paved the way for several other initiatives [36]. Now, there are many countries aiming at 100 thousand WGS along with extensive clinical data to improve precision medicine initiatives by providing both reference datasets and research substrates for the discovery of novel genes and loci associated with traits.

It is important to mention the advantages and disadvantages of WGS over other methods. As compared to WES, sequencing the entire genome allows interrogation of both common and rare variants within and outside coding regions. The high-density microarrays are useful when researchers are agnostically detecting association signals across the genome (when performing GWAS), and more often than not, signals fall within intergenic or intronic regions. If the association is truly a proxy of causal variants nearby, WGS would be useful in both steps, allowing fine-mapping to pinpoint candidates of causality within variants of lower frequency. WGS uniformity in horizontal coverage allows an improvement in phasing (or haplotyping) estimates (the attribution of the relative position of two or more alleles in *cis* or *trans* configurations). By chance, paired-end reads can harbor informative variants that aid algorithms to keep track of the phase, which can be useful in classification pathogenicity of variants in recessively inherited disorders. Phasing software can take advantage of this information together with estimated haplotype frequencies to infer haplotypes throughout the genome. The alternative to that is applying the gold-standard phasing procedure, which analyzes trios and duos; however, this strategy is often disregarded in favor of sampling more unrelated individuals. In addition, both properties of long-range horizontal coverage combined with uniform vertical coverage facilitate detection, mapping, and description of SVs, which will be covered in the next section.

An important disadvantage of WGS over WES or targeted panels is the cost, not only of equipment and reagents but also of computational resources to process and store generated data. While the wet lab steps (including sequencing runs) of WGS had just breached down the 1000USD barriers, storage alone can represent 5% of this value per year. In addition to that, annotation of noncoding regions is still challenging and the gain in diagnostic capacities from WES to WGS is not yet clear. For research purposes, on the other hand, WGS is an excellent tool that embraces many analytical possibilities [37].

NGS Analyses

The extensive use of NGS-based tests and NGS for research gave rise to a whole community of users, composed of wet-lab researchers and technicians, bioinformaticians, programmers, and analysts. Two interesting things came as a result of this: standardized protocols and recommended guidelines were developed and improved over time, building confidence and reproducibility of results; and a vast universe of alternative methods were tested allowing researchers to apply NGS in several different manners. This section will focus on the basic pipeline and comment on workflows that support the most common applications.

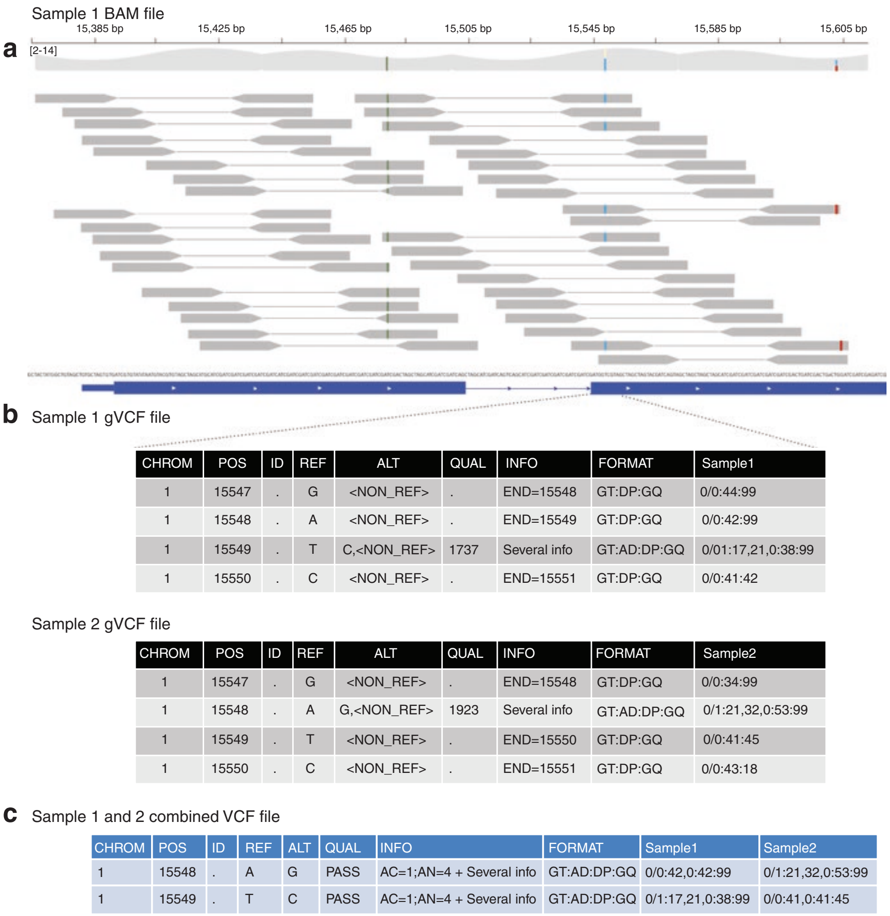
As previously described, the sequencer captures the position and intensities of clusters across flow cells, cycle after cycle, with a camera-like sensor. To reduce the volume of data generated at this process, a Real-Time Analyzer software (Illumina proprietary resource) converts data to BCL format, a binary file that contains raw information on each cluster output. This file must be exported to a server where the primary bioinformatics pipeline will take place. The first step involves demultiplexing and conversion of BCL to FASTQ files, a text-based file that stores the sequence ID (including the cluster position, useful for paired-end sequencing), the sequence itself, and per-base sequencing quality.

At this point, mapping can proceed using alignment software. Two strategies can take place: *de novo* assembly or alignment to a genome reference. Keep in mind that the original Human Genome Project had precisely this challenge (besides generating all raw data): assembly is computationally costly since all reads must query themselves to build up contigs based on overlapping stretches. Since then, updates on the human reference genome were made, and interesting discussions on how a reference should be represented to account for diversity became common. Alignment algorithms can be tuned in several parameters: if they are too strict, reads containing true alternate alleles might not be considered; if they are too permissive, mapping quality will decrease, since reads will align to many loci. Also, indels and larger

SVs tend to penalize alignments and depending on the size of the variant, the read length itself, and the genomic context, some of those variants will not be aligned and, therefore, will not be called. BWA-MEM is the more commonly used free aligner and outputs a raw BAM file (the binary version of a SAM file) which provides information on the position of the reads relative to the reference genome of choice and mapping quality. Then, a few intermediate steps that involve marking and removing PCR duplicates, local realignment for improved *indel* detection, and recalibrating base scores to the local and overall sample context are performed to obtain an analysis-ready BAM file. BAM files are ready for visualization by a genome browser such as IGV, providing an image of piled-up reads aligned to a reference genome sequence (Fig. 3.3a). Base mismatches of the aligned data could only mean three things: true variants, sequencing errors or alignment errors. The following step is to integrate the mismatches across the reads, effectively calling variants. HaplotypeCaller is the variant calling tool recommended by the Genome Analysis Toolkit (GATK), a recommended general protocol provided and maintained by the Broad Institute of Harvard and MIT and used worldwide [38]. It outputs a (very large) file, named gVCF, that has the same format of a standard VCF, except it contains the genotypes of all called positions, whether it is a variant call or not (Fig. 3.3b).

The gVCF can also block information whenever a sample has reference alleles for consecutive stretches (indicating the end of each block in the INFO column), as well as an indication of the spot for a non-reference allele. The FORMAT column will contain guidance to read the genotypes, which generally contain the inferred genotype itself referring to REF and ALT status, depth of coverage for each allele (AD), depth of coverage at the site (DP) and genotype quality score (GQ). Some files can contain a strand-specific allele depth, which can be useful to evaluate strand bias. The ID column represents the only place with “outside” information (an annotation, by definition), and should be completed with rsID from dbSNP. INFO column is usually populated with several quality statistics: in gVCF refers to the site and individual genotypes, but in combined VCF may include the overall site quality and other metrics such as allele count (AC), allele number (AN), inbreeding coefficient, and Hardy-Weinberg statistics. The same for QUAL, which individually represents the site and genotype quality but in conjunction with other samples, provides a flag with confidence levels for the site (Fig. 3.3c).

Next, this individual gVCFs can be combined with other samples in a joint cohort to a joint-call step that will result in a standard VCF file containing only positions in which there is at least one alternate allele. The header of VCF files can store a number of meta-information, including the description of the entries present in the body of the VCF file and the commands used to obtain the VCF. The steps presented here compose the pipeline recommended by the GATK Best Practices to identify germline short-range variants, which is used by most bioinformaticians and very appraised by the scientific community. The GATK Best Practices validated pipelines with recommended software, quality parameters, and continued improvements for different types of variants.



Sample 1 gVCF file

CHROM	POS	ID	REF	ALT	QUAL	INFO	FORMAT	Sample1
1	15547	.	G	<NON_REF>	.	END=15548	GT:DP:GQ	0/0:44:99
1	15548	.	A	<NON_REF>	.	END=15549	GT:DP:GQ	0/0:42:99
1	15549	.	T	C,<NON_REF>	1737	Several info	GT:AD:DP:GQ	0/01:17,21,0:38:99
1	15550	.	C	<NON_REF>	.	END=15551	GT:DP:GQ	0/0:41:42

Sample 2 gVCF file

CHROM	POS	ID	REF	ALT	QUAL	INFO	FORMAT	Sample2
1	15547	.	G	<NON_REF>	.	END=15548	GT:DP:GQ	0/0:34:99
1	15548	.	A	G,<NON_REF>	1923	Several info	GT:AD:DP:GQ	0/1:21,32,0:53:99
1	15549	.	T	<NON_REF>	.	END=15550	GT:DP:GQ	0/0:41:45
1	15550	.	C	<NON_REF>	.	END=15551	GT:DP:GQ	0/0:43:18

Sample 1 and 2 combined VCF file

CHROM	POS	ID	REF	ALT	QUAL	INFO	FORMAT	Sample1	Sample2
1	15548	.	A	G	PASS	AC=1;AN=4 + Several info	GT:AD:DP:GQ	0/0:42,0:42:99	0/1:21,32,0:53:99
1	15549	.	T	C	PASS	AC=1;AN=4 + Several info	GT:AD:DP:GQ	0/1:17,21,0:38:99	0/0:41,0:41:45

Fig. 3.3 Schematic representation of BAM file from one individual (**a**), gVCF files from two individuals (**b**), and combined VCF file with two individuals (**c**). In the BAM file (**a**) we can see an example of paired-end reads containing variants in colors (all grey portions of the reads are matching the reference). Genomic position is represented as a ruler above, and depth of coverage as the light grey graph below. Below the reads, the reference genome is shown along with a basic gene/transcript annotation in blue (in this case, representing the two first exons, where the first has a 5'UTR portion and direction of gene in the genome). Note that paired-end brought evidence of phasing between the first green variant (in homozygous state) with the second blue variant (in heterozygous state), which in turn is also in cis with the third red variant. gVCF files contain all called positions (**b**), and variant based quality (which is also included in the INFO field, along with several other metrics of sequencing, alignment, base calling, and variant calling). VCF file summarizes positions with variants (**c**) and includes site quality and information (now as an aggregate of all individuals included in this combined file). This is ready for annotation and downstream analyses

The steps presented here compose the pipeline recommended by the Genome Analysis Toolkit (GATK) Best Practices to identify germline short-range variants, which is used by most bioinformaticians and very appraised by the scientific community. The GATK Best Practices is a product of the collaborative effort provided by the Broad Institute (Boston, MA) which thoroughly validated pipelines and provided a workflow with recommended software, quality parameters, and continued improvements for different types of variants.

Either an individual VCF or a combined VCF (with multiple individuals) can now be annotated to provide context to the findings. As mentioned above, annotation is a procedure that will systematically intersect findings with previous knowledge stored in datasets and includes straightforward basic annotations such as rsID, gene, and transcripts. There are several annotations that can be relevant for various analyses such as the frequency of variants across different datasets, the association of genes to disorders, pathogenicity assertions, prediction of deleteriousness by different algorithms, context of protein domains. There are several annotators in use, most freely available such as ANNOVAR [39] and Ensembl Variant Effect Predictor (VEP) [40], but it is common that laboratories add in-house scripts for specific annotations.

In the previous sections, we have stated that NGS-based analyses are not the gold-standard method for the detection of large SVs. One reason is that uniformity of reads (both in vertical and horizontal coverage) is not always predictable and varies within the individual sample and across samples. For instance, although the peaks of depth surrounding an exon in a WES sample can reach over 200x (in a sample aimed at 100x), it is not trivial to infer if that particular exon was better captured than the others or if it represented a duplication. The same goes for heterozygous deletions: a drop in depth of coverage can be caused by a deletion or by a lower capture performance. However, there are several algorithms and workflows that use read-depth measures to successfully detect a high proportion of CNVs from NGS data, most of the time through exome or gene panels. In fact, NGS-based CNV analysis is increasing in both clinical and research contexts as a cost-effective choice to study a broader range of variants [41]. The optimum choice for short-range NGS-based CNV analysis would be to use paired-end deep-coverage (>30x) WGS, which has the main advantage of more coverage uniformity (i.e. less variation of depth along chromosomes and among individuals). This characteristic facilitates the definition of a reference depth to which deviations can be tested and allow the extensive use of read-pair (RP) and split-read (SR) methods [42]. RP detects discordant pairs in which the span and/or the orientation of read-pairs are inconsistent with the expected insert size. If a deletion spans a well-covered region, paired-end reads will align to the boundaries of the deletion, and will appear to be more distant from each other than expected (Fig. 3.4). On the other hand, read-pairs closer than expected indicate insertions. SR identifies split sequence-read signatures breaking the alignment to the reference (gaps indicate deletions and stretches indicate insertions), detecting the precise boundaries of the variation (breakpoints). RP and SR are useful to identify other types of SVs, including inversions and translocations [43].

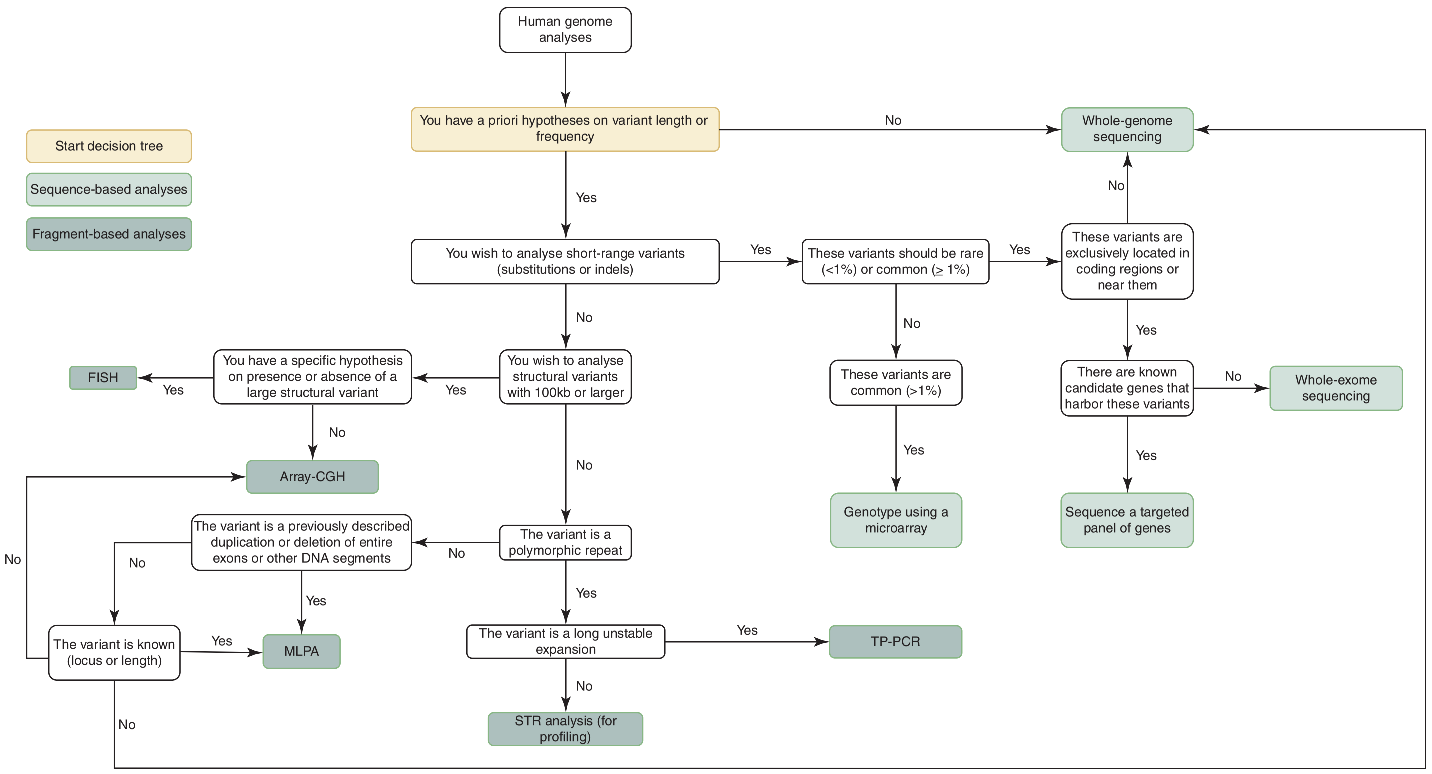

Start decision tree
Sequence-based analyses
Fragment-based analyses
Human genome analyses
You have a priori hypotheses on variant length or frequency
No
Whole-genome sequencing
Yes
You wish to analyse short-range variants (substitutions or indels)
Yes
These variants should be rare (<1%) or common (≥ 1%)
Yes
These variants are exclusively located in coding regions or near them
No
No
These variants are common (>1%)
Yes
Genotype using a microarray
Yes
There are known candidate genes that harbor these variants
No
Whole-exome sequencing
Yes
Sequence a targeted panel of genes
No
You wish to analyse structural variants with 100kb or larger
Yes
You have a specific hypothesis on presence or absence of a large structural variant
Yes
FISH
No
Array-CGH
No
The variant is a polymorphic repeat
No
The variant is a previously described duplication or deletion of entire exons or other DNA segments
Yes
MLPA
No
The variant is known (locus or length)
Yes
No
No
Yes
The variant is a long unstable expansion
Yes
TP-PCR
No
STR analysis (for profiling)

Fig. 3.1 Decision tree of selected methods for genomic analyses, based on variant length, frequency, and constitution. Here, the decision was guided by choosing the current gold standard methods for each aim; however, as discussed along with the text, the choice will ultimately favor the best cost-effective method

3.4 Analysis of Rare and Common Variants to Understand Diseases and Traits

3.4.1 Workflow for Molecular Diagnosis

Molecular diagnosis for patients affected by rare genetic disorders with monogenic patterns of inheritance is straightforward [35]. It begins with a deep clinical evaluation of the patients and family members, which will provide clues for diagnostic hypotheses. Family history of disorders or related phenotypes, occurrence of consanguineous marriages in the family, age of manifestation and clinical progression, and age of parents help the physician narrow down possible candidates. Further access to public databases such as OMIM, Orphanet, and GeneReviews, indicate one or more genes previously associated with the condition or part of the phenotypes that can be prioritized and interrogated with methods described in this chapter. Choosing the best method to start the diagnostic investigation is not trivial: previous knowledge on the disease and genes is critical for establishing a rational stepwise set of tests. Once the test is performed, a complete pipeline for analysis, including access to databases of variants and related literature, will provide a report to be returned in genetic counseling consultation. In the case the test is NGS-based, such as targeted panels or WES, there are general recommendations provided by the American College of Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) that support the workflow, pathogenicity classification, and handling of secondary findings [19]. There are several ethical concerns to this process, which are not the scope of this chapter.

There are, however, concerns regarding the clinical level of evidence and penetrance of variants that must be addressed. Monogenic disorders are generally caused by one or two variants of large effect size, meaning that the presence of such variants greatly increases the risk of manifestation (up to complete penetrance in some cases such as Huntington's disease). Most common disorders, however, have multifactorial etiologies, with an environmental component and a genetic component. Given that the genetic component is usually polygenic, causative variants have, individually, low to moderate effect sizes and are distributed across dozens to hundreds of loci. In research, large GWAS efforts identified many loci associated with multifactorial traits, indeed improving the knowledge on the architecture of such traits and unraveling part of molecular and cellular mechanisms involved in these conditions. Some alleles in genes such as *APOE* were robustly associated with Alzheimer's disease and cognitive impairment with a relatively high odds ratio (3–15, depending on the study and zygosity), but as any single susceptibility allele of a multifactorial disorder, it explains only partially the phenotypic variability [44]. Even though companies offer tests and reports with variants of reduced penetrance in direct-to-consumer tests, the clinical validity of such associations is still under discussion by scientific and medical societies.

3.4.2 Rare-Variant Association Testing

Although successfully discovering tens of thousands of variants robustly associated with diverse traits and better understanding their genetic architecture, GWAS hits explain only a small proportion of phenotypic variability (a problem referred to as the missing heritability). GWAS is, by design, focused on common variants (usually defined as variants with minor allele frequency > 0.01) obtained from genotyping are interrogated using microarrays. Part of the missing heritability arises because many common variants have very small effect sizes that could be detected only with increasing sample sizes. The ever-increasing sample sizes of recent GWAS and meta-analyses intend to address this problem. Another part of the missing heritability arises from not considering rare variants, which were shown to collectively reach significant effect sizes. However, even if GWAS included rare variants in its analysis, detecting association with standard GWAS protocols would require enormous sample sizes. To address this issue, methods for collapsing rare variants per gene, per genomic region, or per pathway have been developed and improved overtime. The rationale is quite straightforward: case and control groups of individuals are sequenced (ideally WES or WGS) and rare variants are computed per group, within candidate genes, regions, or pathways, or alternatively, multiple combinations of genes are tested. Variant annotations can be used in weighing each aggregate, and some tests are prepared to combine common variants as well. Many publicly available algorithms perform these tests, also known as burden and nonburden tests, that can well complement GWAS or be used when larger sample sizes are not available [45, 46].

3.4.3 Polygenic Risk Scores

GWAS usually identifies individual association signals for each variant and strict thresholds are applied to ensure the exclusion of false-positive results. However, as mentioned, there are many loci truly associated with the traits that do not reach statistical significance due to several reasons: reduced effect size, low frequency, population-specific linkage disequilibrium patterns, and epistasis with more loci. With the recent availability of very large cohorts with genotyping data and comprehensive phenotypic information, such as the UK BioBank, testing combinations of variants based on GWAS summary statistics (effect sizes measured as odds ratio or regression betas) could be performed [47]. Using a protocol for reducing interdependence of variant's signal by pruning blocks under linkage disequilibrium and thresholding *p*-values, many researchers are exploring UK BioBank's large sample sizes (about 500 thousand genotyped individuals) to identify combinations of variants that successfully stratify individuals by disease risk (or trait levels). These profiles can be interrogated in a validation set, providing a distribution of a polygenic risk score (PRS). Both the scientific communities and medical societies are enthusiastic about the application of PRS in several traits and disorders, since these profiles would eventually anticipate clinical interventions for individuals at higher risks, at ultimately low costs (microarrays). The full extent of rare variant contributions to PRS is yet to be elucidated. However, there are already studies on breast cancer and hypercholesterolemia showing the combined effect of different PRS risks and carrier status of high effect size monogenic variants, providing a good perspective on the clinical applications of PRS and that WGS might be the ultimate test to embrace all dimensions [48].

One important drawback, currently under discussion, is that source samples used in GWAS are still biased towards Europeans, and PRS transferability to other populations is challenging, with significant reductions of predictive power (See Chap. 11). Holding the same caution alert on direct-to-consumer testing, it is noteworthy that admixed individuals or individuals from diverse populations different from the original sourced in large GWAS might not benefit from such PRS-based tests and may receive reports with reduced clinical validity [49].

3.5 Perspectives

A brief glimpse of the rapid development and improvements in genomic analysis methods was presented in the previous sections. The consolidation of NGS as the gold-standard sequencing method does not mean that all challenges imposed by genomic complexity have been fully addressed. In this section, our intention is to introduce selected methods under implementation that are likely to complement or eventually replace the current protocols.

3.5.1 Cell-Free DNA

Cell-free DNA (cfDNA) refers to any degraded DNA fragment present in the circulation and other biological fluids. They were first detected in 1948 and since then many studies investigating their possible association with different diseases were conducted. Currently, detection and analyses of cfDNA originated from tumors and from the fetus in a pregnant woman are widely used. The rationale is to detect somatic mutations from tumor and either chromosomal imbalances (such as aneuploidies) or *de novo* mutations from the fetus (absent from mother and father). Plasma cfDNA concentration is usually low, and DNA is very fragmented, so the depth of coverage usually is aimed higher and paired comparisons are performed: plasma from patient vs buffy coat (blood fraction containing mononucleated cells) from the same patient, or plasma from mother vs buffy coat or saliva from mother [50].

The investigation of cfDNA from tumors is known as liquid biopsy, a noninvasive procedure that allows routine clinical screening to detect resistance mechanisms to inform treatment, and to monitor the response to treatment and residual disease. Besides, evaluating the use of liquid biopsy as a clinical tool for early cancer detection is currently an active area of research.

Another widely application of this type of sample is to obtain cfDNA from the fetus and screen for the most common aneuploidies through NGS sequencing. This test, known as noninvasive prenatal testing (NIPT) or noninvasive prenatal screening (NIPS), has the advantage of being noninvasive, in contrast to invasive procedures to obtain fetal DNA which brings risks to pregnancy, and of being more accurate. The exam can be performed from the ninth week of gestation and although it is highly accurate, it is important to keep in mind that the accuracy is not high enough to be considered a diagnostic test. Although the focus is on detecting common aneuploidies, other important known SVs, particularly microdeletion syndromes can be investigated. Besides, methods to investigate monogenic diseases are currently being developed and validated [51].

3.5.2 Long-Read Sequencing

NGS-based methods use relatively short-reads that challenge the determination of several types of genomic variation, such as SVs, pseudogenes, and highly similar genes, highly repetitive regions (including disease-related repeat expansions), and highly diverse haplotypes (such as HLA regions). Besides, short-reads complicate inference of phasing information (for compound heterozygosity determination), particularly for very rare variants and singletons. Some methodological improvements address these issues by modifying NGS libraries preparation using mate-pair (Illumina), linked-reads (10X), and Hi-C assays, that capture three-dimensional chromatin conformation and provide evidence on structural interactions. However, all-new technologies arose in the past decade to fill this gap, with continuous fragments sequenced from few kb to megabases (named long-reads), directly obtained from native DNA. We will briefly explore two main platforms for long-read sequencing, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) [52].

PacBio developed a protocol named SMRT sequencing (*single-molecule, real-time*) which uses kilobase-long fragments (up to more than 100 kb) with adapters connected to their ends forming hairpins (leaving the double-strand capped, able to circularize in complementary single strands). This structure is then assembled with DNA polymerase for loading into a SMRT Cell, a much reduced flow cell with a nanophotonic surface containing zero-mode waveguides (ZMW), which works as chamber reactions and photons. The incorporation of fluorescently labeled dNTPs excites the dye and a camera sensor, which detects the incorporation in real time for each ZMW. Fluorophore is removed before each light pulse to prevent spectral interference. Each forward and reverse strands keep circulating for some rounds, providing a measure of depth of coverage. Median read length in PacBio is around 10–60 kb and although read accuracy averages 90%, there are high fidelity protocols that reach over 99%, with a 10–30 Gb throughput per flow cell. In this protocol, cost per Gb reaches about 86USD (over twice Illumina's NovaSeq current pricing).

ONT begins with a linear DNA fragmented in long stretches (from 1 kb to a few megabases long) that are in double-strand form and are ligated to a sequence adapter attachable to a motor protein. The flow cells are composed of membranes embedded with thousands of engineered nanopores, with which motor proteins will interact and unwind DNA into the pores, transiently disrupting the electric current of the membrane. The current changes are base-pair specific and since the rate of translocation is controlled, there is enough resolution for detection of individual base pairs and homopolymers. Detection occurs in real-time and although there are larger platforms for loading multiple flow cells, ONT became famous for providing the smallest sequencer ever created to this day (MinION), with the size of a smartphone and the ability to be transported. The ability to consistently sequence reads of dozens of kb and ease of transport allowed MinION to become very popular among microbiologists during field trips, and more recently, to sequence the entire viral genomes of Zika and SARS-CoV-2, in many locations [53, 54]. With variable accuracy of 87–98% depending on the platform, ONT reads range from 10 to 200 kb (but reads over 1 Mb were obtained and replicated), with outputs from 2 to 100Gb per flow cell and costs approaching Illumina's NovaSeq.

Several important accomplishments were already achieved by long-read sequencing, including the ability to distinguish modified bases such as methylation state of cytosines, relevant in epigenomic studies (See Chap. 4). In addition to that, native RNA sequencing has been reported with ONT and replicated and promises an interesting future in identifying full-length transcript isoforms. Modification in RNA bases (epitranscriptomics) is also being explored using ONT. Accuracy of long-read sequencing is still behind as compared to NGS, but combining both methods have demonstrated a significant gain in de novo genome assembly and confirmation of structural variation. There are novel methods for target enrichment of DNA loci using CRISPR-Cas9 that, in theory, improve accuracy of long-read sequencing by increasing depth of coverage. Adoption of long-read sequencing may improve our ability to detect, catalogue, and interpret haplotypes directly inferred from genomes and transcriptomes.

3.5.3 Omics Integration

The studies on genetic variation are often descriptive, as providing evidence on the consequences of this type of findings is challenging. While there is not a truly systematic strategy of defining functional consequences to molecules, cells, tissues, organs, and clinical manifestations of all variants found in one individual, a group of patients, or a population, there are orthogonal methods that help drawing a larger picture.

The ENCODE Project Consortium has the objective of deeply annotating DNA elements of the genome by integrating research groups and methods to describe and validate regions of the genome that interact with transcription factors, chromatin structure, and methylation sites [15]. The Genotype-Tissue Expression (GTEx) project, on the other hand, intended the creation of a resource on gene expression and its regulation in dozens of human tissues, providing a full description of variation, expression conditions, and transcriptional outcomes [55]. More recently, the Human Cell Atlas project was launched to integrate research groups involved in cellular models, including genomics, transcriptomics, proteomics, and metabolomics. Novel technologies such as single-cell transcriptomics, allows a deep description of cellular states under different conditions and mapping signatures involved in pathology can give insight in gene function [56, 57].

In addition, large initiatives that intend to collect and follow up clinical data and other traits in a population-level scale along with biological sampling, as an example of the UK BioBank, are already contributing for our understanding of the association between genomic variability and outcomes. Combining data from hundreds of thousands of individuals improves detection of small effect variants and polygenic profiles. Several biobanks also include other levels of biomedically relevant experiments such as RNA-Seq, epigenomics, proteomics, and metabolomics, all of which can be integrated using both agnostic approaches such as deep learning or candidate-driven by piling up individual-level information (“thick-data”) [58].

In this chapter, we presented how genomic methods have constantly been evolving over the past decades, and with all the new technologies and the enormous population samples being analyzed, we can expect that the coming years will continue to bring significant advances to genomic science, ultimately making precision medicine a reality in clinical routine.

References

1. Lejeune J, Gautier M, Turpin R. Study of somatic chromosomes from 9 mongoloid children. C R Hebd Seances Acad Sci. 1959;248(11):1721–2.
2. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737–8. https://doi.org/10.1038/171737a0.
3. O'Brien SJ, MacIntyre RJ. An analysis of gene-enzyme variability in natural populations of Drosophila melanogaster and *D. simulans*. Am Nat. 1969;103(930):97–113.
4. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7. https://doi.org/10.1073/pnas.74.12.5463.
5. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 1986;51:263–73. https://doi.org/10.1101/sqb.1986.051.01.032.
6. Cook-Deegan RM. The Alta summit, December 1984. Genomics. 1989;5(3):661–3. https://doi.org/10.1016/0888-7543(89)90042-6.

7. Eichler EE. Genetic variation, comparative genomics, and the diagnosis of disease. N Engl J Med. 2019;381(1):64–74. https://doi.org/10.1056/NEJMra1809315.
8. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, Buyske S, Matise TC, Muzny DM, Zody MC, Lander ES, Dutcher SK, Stitziel NO, Hall IM. Mapping and characterization of structural variation in 17,795 human genomes. Nature. 2020;583(7814):83–9. https://doi.org/10.1038/s41586-020-2371-0.
9. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D, Anantharaman T, Audano PA, Brand H, Cantsilieris S, Cao H, Cerveira E, Chen C, Chen X, Chin CS, Chong Z, Chuang NT, Lambert CC, Church DM, Clarke L, Farrell A, Flores J, Galeev T, Gorkin DU, Gujral M, Guryev V, Heaton WH, Korlach J, Kumar S, Kwon JY, Lam ET, Lee JE, Lee J, Lee WP, Lee SP, Li S, Marks P, Viaud-Martinez K, Meiers S, Munson KM, Navarro FCP, Nelson BJ, Nodzak C, Noor A, Kyriazopoulou-Panagiotopoulou S, Pang AWC, Qiu Y, Rosanio G, Ryan M, Stutz A, Spierings DCJ, Ward A, Welch AE, Xiao M, Xu W, Zhang C, Zhu Q, Zheng-Bradley X, Lowy E, Yakneen S, McCarroll S, Jun G, Ding L, Koh CL, Ren B, Flicek P, Chen K, Gerstein MB, Kwok PY, Lansdorp PM, Marth GT, Sebat J, Shi X, Bashir A, Ye K, Devine SE, Talkowski ME, Mills RE, Marschall T, Korbel JO, Eichler EE, Lee C. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1784. https://doi.org/10.1038/s41467-018-08148-z.
10. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee S-b, Tian X, Browning BL, Das S, Emde A-K, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Wong Q, Aguet F, Albert C, Alonso A, Ardlie KG, Aslibekyan S, Auer PL, Barnard J, Barr RG, Becker LC, Beer RL, Benjamin EJ, Bielak LF, Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chen Y-DI, Cho MH, Choi SH, Chung MK, Clish CB, Correa A, Curran JE, Custer B, Darbar D, Daya M, Andrade Md, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Fatkin D, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Köttgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin K-H, Liu C, Loos RJF, Garman L, Gerszten R, Lubitz SA, Lunetta KL, Mak ACY, Manichaikul A, Manning AK, Mathias RA, McManus DD, McGarvey ST, Meigs JB, Meyers DA, Mikulla JL, Minear MA, Mitchell B, Mohanty S, Montasser ME, Montgomery C, Morrison AC, Murabito JM, Natale A, Natarajan P, Nelson SC, North KE, O'Connell JR, Palmer ND, Pankratz N, Peloso GM, Peyser PA, Post WS, Psaty BM, Rao DC, Redline S, Reiner AP, Roden D, Rotter JI, Ruczinski I, Sarnowski C, Schoenherr S, Seo J-S, Seshadri S, Sheehan VA, Shoemaker MB, Smith AV, Smith NL, Smith JA, Sotoodehnia N, Stilp AM, Tang W, Taylor KD, Telen M, Thornton TA, Tracy RP, Berg DJVD, Vasan RS, Viaud-Martinez KA, Vrieze S, Weeks DE, Weir BS, Weiss ST, Weng L-C, Willer CJ, Zhang Y, Zhao X, Arnett DK, Ashley-Koch AE, Barnes KC, Boerwinkle E, Gabriel S, Gibbs R, Rice KM, Rich SS, Silverman E, Qasba P, Gan W, Papanicolaou GJ, Nickerson DA, Browning SR, Zody MC, Zöllner S, Wilson JG, Cupples LA, Laurie CC, Jaquish CE, Hernandez RD, O'Connor TD, Abecasis GR (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. bioRxiv.
11. Lloyd JP, Tsai ZT-Y, Sowers RP, Panchy NL, Shiu S-H, Gojobori J. A model-based approach for identifying functional intergenic transcribed regions and noncoding RNAs. Mol Biol Evol. 2018;35(6):1422–36. https://doi.org/10.1093/molbev/msy035.
12. Gonzalez-Sandoval A, Gasser SM. On TADs and LADs: spatial control over gene expression. Trends Genet. 2016;32(8):485–95. https://doi.org/10.1016/j.tig.2016.05.004.
13. Melo US, Schöpflin R, Acuna-Hidalgo R, Mensah MA, Fischer-Zirnsak B, Holtgrewe M, Klever M-K, Türkmen S, Heinrich V, Pluym ID, Matoso E, Bernardo de Sousa S, Louro P, Hülsemann W, Cohen M, Dufke A, Latos-Bieleńska A, Vingron M, Kalscheuer V,

Quintero-Rivera F, Spielmann M, Mundlos S. Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases. Am J Hum Genet. 2020;106(6):872–84. https://doi.org/10.1016/j.ajhg.2020.04.016.

14. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5(4):eaaw1668. https://doi.org/10.1126/sciadv.aaw1668.
15. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. https://doi.org/10.1038/nature11247.
16. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Genome Aggregation Database C, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7.
17. Antonarakis SE. Carrier screening for recessive disorders. Nat Rev Genet. 2019;20(9):549–61. https://doi.org/10.1038/s41576-019-0134-2.
18. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation C. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91. https://doi.org/10.1038/nature19057.
19. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, Committee ALQA. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. Genetics in Medicine: Official Journal of the American College of Medical Genetics. 2015;17(5):405–24. https://doi.org/10.1038/gim.2015.30.
20. Van Prooijen-Knegt AC, Van Hoek JFM, Bauman JGJ, Van Duijn P, Wool IG, Van der Ploeg M. In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. Exp Cell Res. 1982;141(2):397–407. https://doi.org/10.1016/0014-4827(82)90228-2.
21. The BAC Resource Consortium, Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M, Conroy J, Kasprzyk A, Massa H, Yonescu R, Sait S, Thoreen C, Snijders A, Lemyre E, Bailey JA, Bruzel A, Burrill WD, Clegg SM, Collins S, Dhami P, Friedman C, Han CS, Herrick S, Lee J, Ligon AH, Lowry S, Morley M, Narasimhan S, Osoegawa K, Peng Z, Plajzer-Frick I, Quade BJ, Scott D, Sirotkin K, Thorpe AA, Gray JW, Hudson J, Pinkel D, Ried T, Rowen L, Shen-Ong GL, Strausberg RL, Birney E, Callen DF, Cheng JF, Cox DR, Doggett NA, Carter NP, Eichler EE, Haussler D, Korenberg JR, Morton CC, Albertson D, Schuler G, de Jong PJ, Trask BJ. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature. 2001;409(6822):953–8. https://doi.org/10.1038/35057192.

22. Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, Garini Y, Ried T. Multicolor spectral karyotyping of human chromosomes. Science. 1996;273(5274):494–7. https://doi.org/10.1126/science.273.5274.494.
23. Speicher MR, Ballard SG, Ward DC. Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat Genet. 1996;12(4):368–75. https://doi.org/10.1038/ng0496-368.
24. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258(5083):818–21. https://doi.org/10.1126/science.1359641.
25. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998;20(2):207–11. https://doi.org/10.1038/2524.
26. Shaffer LG, Bejjani BA, Torchia B, Kirkpatrick S, Coppinger J, Ballif BC. The identification of microdeletion syndromes and other chromosome abnormalities: cytogenetic methods of the past, new technologies for the future. Am J Med Genet C: Semin Med Genet. 2007;145C(4):335–45. https://doi.org/10.1002/ajmg.c.30152.
27. Schouten JP. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic acids Res. 2002;30(12):57e–57. https://doi.org/10.1093/nar/gnf056.
28. Warner JP, Barron LH, Goudie D, Kelly K, Dow D, Fitzpatrick DR, Brock DJ. A general method for the detection of large CAG repeat expansions by fluorescent PCR. J Med Genet. 1996;33(12):1022–6. https://doi.org/10.1136/jmg.33.12.1022.
29. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345–53. https://doi.org/10.1038/nature24286.
30. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol. 2013; https://doi.org/10.1002/0471142727.mb2201s101.
31. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20(8):467–84. https://doi.org/10.1038/s41576-019-0127-1.
32. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:1–11. https://doi.org/10.1155/2012/251364.
33. Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, Roychowdhury S. Evaluation of hybridization capture versus amplicon-based methods for whole-Exome sequencing. Hum Mutat. 2015;36(9):903–14. https://doi.org/10.1002/humu.22825.
34. Weiss MM, Van der Zwaag B, Jongbloed JDH, Vogel MJ, Brüggenwirth HT, Lekanne Deprez RH, Mook O, Ruivenkamp CAL, van Slegtenhorst MA, van den Wijngaard A, Waisfisz Q, Nelen MR, van der Stoep N. Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a National Collaborative Study of Dutch genome diagnostic laboratories. Hum Mutat. 2013;34(10):1313–21. https://doi.org/10.1002/humu.22368.
35. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;19:253. https://doi.org/10.1038/nrg.2017.116.
36. The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
37. Koboldt Daniel C, Steinberg Karyn M, Larson David E, Wilson Richard K, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155(1):27–38. https://doi.org/10.1016/j.cell.2013.09.006.

38. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. https://doi.org/10.1101/gr.107524.110.
39. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. https://doi.org/10.1093/nar/gkq603.
40. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome Biol. 2016;17(1):122. https://doi.org/10.1186/s13059-016-0974-4.
41. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Front Bioeng Biotechnol. 2015;3:92. https://doi.org/10.3389/fbioe.2015.00092.
42. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333–9. https://doi.org/10.1093/bioinformatics/bts378.
43. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, Raca G, Ritter DI, South ST, Thorland EC, Pineda-Alvarez D, Aradhya S, Martin CL. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the clinical genome resource (ClinGen). Genet Med. 2019;22(2):245–57. https://doi.org/10.1038/s41436-019-0686-8.
44. Ridge PG, Hoyt KB, Boehme K, Mukherjee S, Crane PK, Haines JL, Mayeux R, Farrer LA, Pericak-Vance MA, Schellenberg GD, Kauwe JSK, Adams PM, Albert MS, Albin RL, Apostolova LG, Arnold SE, Asthana S, Atwood CS, Baldwin CT, Barber RC, Barmada MM, Barnes LL, Barral S, Beach TG, Becker JT, Beecham GW, Beekly D, Bennett DA, Bigio EH, Bird TD, Blacker D, Boeve BF, Bowen JD, Boxer A, Burke JR, Burns JM, Buxbaum JD, Cairns NJ, Cantwell LB, Cao C, Carlson CS, Carlsson CM, Carney RM, Carrasquillo MM, Carroll SL, Chui HC, Clark DG, Corneveaux J, Crane PK, Cribbs DH, Crocco EA, Cruchaga C, De Jager PL, DeCarli C, Demirci FY, Dick M, Dickson DW, Doody RS, Duara R, Ertekin-Taner N, Evans DA, Faber KM, Fairchild TJ, Fallon KB, Fardo DW, Farlow MR, Ferris S, Foroud TM, Frosch MP, Galasko DR, Gearing M, Geschwind DH, Ghetti B, Gilbert JR, Goate AM, Graff-Radford NR, Green RC, Growdon JH, Hakonarson H, Hamilton RL, Hamilton-Nelson KL, Hardy J, Harrell LE, Honig LS, Huebinger RM, Huentelman MJ, Hulette CM, Hyman BT, Jarvik GP, Jicha GA, Jin L-W, Jun G, Kamboh MI, Karydas A, Katz MJ, Kauwe JSK, Kaye JA, Kim R, Kowall NW, Kramer JH, Kukull WA, Kunkle BW, LaFerla FM, Lah JJ, Larson EB, Leverenz JB, Levey AI, Li G, Lieberman AP, Lin C-F, Lipton RB, Lopez OL, Lunetta KL, Lyketsos CG, Mack WJ, Marson DC, Martin ER, Martiniuk F, Mash DC, Masliah E, McCormick WC, McCurry SM, McDavid AN, McKee AC, Mesulam M, Miller BL, Miller CA, Miller JW, Montine TJ, Morris JC, Mukherjee S, Murrell JR, Myers AJ, Naj AC, O'Bryant S, Olichney JM, Pankratz VS, Parisi JE, Partch A, Paulson HL, Perry W, Peskind E, Petersen RC, Pierce A, Poon WW, Potter H, Quinn JF, Raj A, Raskind M, Reiman EM, Reisberg B, Reisch JS, Reitz C, Ringman JM, Roberson ED, Rogaeva E, Rosen HJ, Rosenberg RN, Royall DR, Sager MA, Sano M, Saykin AJ, Schneider JA, Schneider LS, Seeley WW, Smith AG, Sonnen JA, Spina S, St George-Hyslop P, Stern RA, Swerdlow RH, Tanzi RE, Thornton-Wells TA, Trojanowski JQ, Troncoso JC, Tsuang DW, Valladares O, Van Deerlin VM, Van Eldik LJ, Vardarajan BN, Vinters HV, Vonsattel JP, Wang L-S, Weintraub S, Welsh-Bohmer KA, Wendland JR, Wilhelmsen KC, Williamson J, Wingo TS, Winslow AR, Wishnek S, Woltjer RL, Wright CB, Wu C-K, Younkin SG, Yu C-E, Yu L. Assessment of the genetic variance of late-onset Alzheimer's disease. Neurobiol Aging. 2016;41:200.e213–20. https://doi.org/10.1016/j.neurobiolaging.2016.02.024.

45. Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. Phenotypic extremes in rare variant study designs. Eur J Hum Genet. 2015;24(6):924–30. https://doi.org/10.1038/ejhg.2015.197.
46. Sazonovs A, Barrett JC. Rare-variant studies to complement genome-wide association studies. Annu Rev Genomics Hum Genet. 2018;19(1):97–112. https://doi.org/10.1146/annurev-genom-083117-021641.
47. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219–24. https://doi.org/10.1038/s41588-018-0183-z.
48. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, Lai C, Brockman D, Philippakis A, Ellinor PT, Cassa CA, Lebo M, Ng K, Lander ES, Zhou AY, Kathiresan S, Khera AV. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun. 2020;11(1):3635. https://doi.org/10.1038/s41467-020-17374-3.
49. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. Human demographic history impacts genetic risk prediction across diverse populations. Am J Hum Genet. 2017;100(4):635–49. https://doi.org/10.1016/j.ajhg.2017.03.004.
50. Bronkhorst AJ, Ungerer V, Holdenrieder S. The emerging role of cell-free DNA as a molecular marker for cancer management. Biomol Detect Quantif. 2019;17:100087. https://doi.org/10.1016/j.bdq.2019.100087.
51. Breveglieri G, D'Aversa E, Finotti A, Borgatti M. Non-invasive prenatal testing using fetal DNA. Mol Diagn Ther. 2019;23(2):291–9. https://doi.org/10.1007/s40291-019-00385-2.
52. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21(10):597–614. https://doi.org/10.1038/s41576-020-0236-x.
53. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, Mellan TA, du Plessis L, Pereira RHM, Sales FCS, Manuli ER, Thézé J, Almeida L, Menezes MT, Voloch CM, Fumagalli MJ, Coletti TM, da Silva CAM, Ramundo MS, Amorim MR, Hoeltgebaum HH, Mishra S, Gill MS, Carvalho LM, Buss LF, Prete CA, Ashworth J, Nakaya HI, Peixoto PS, Brady OJ, Nicholls SM, Tanuri A, Rossi ÁD, Braga CKV, Gerber AL, de C. Guimarães AP, Gaburo N, Alencar CS, ACS F, Lima CX, Levi JE, Granato C, Ferreira GM, Francisco RS, Granja F, Garcia MT, Moretti ML, Perroud MW, TMPP C, Lazari CS, Hill SC, de Souza Santos AA, Simeoni CL, Forato J, Sposito AC, Schreiber AZ, MNN S, de Sá CZ, Souza RP, Resende-Moreira LC, Teixeira MM, Hubner J, PAF L, Moreira RG, Nogueira ML, Ferguson NM, Costa SF, Proenca-Modena JL, ATR V, Bhatt S, Lemey P, Wu C-H, Rambaut A, Loman NJ, Aguiar RS, Pybus OG, Sabino EC, Faria NR. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020;369(6508):1255–60. https://doi.org/10.1126/science.abd2161.
54. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, Burton DR, Lewis-Ximenez LL, de Jesus JG, Giovanetti M, Hill SC, Black A, Bedford T, Carroll MW, Nunes M, Alcantara LC, Sabino EC, Baylis SA, Faria NR, Loose M, Simpson JT, Pybus OG, Andersen KG, Loman NJ. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12(6):1261–76. https://doi.org/10.1038/nprot.2017.066.
55. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S,

McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5. https://doi.org/10.1038/ng.2653.

56. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, Zhou Y, Ye F, Jiang M, Wu J, Xiao Y, Jia X, Zhang T, Ma X, Zhang Q, Bai X, Lai S, Yu C, Zhu L, Lin R, Gao Y, Wang M, Wu Y, Zhang J, Zhan R, Zhu S, Hu H, Wang C, Chen M, Huang H, Liang T, Chen J, Wang W, Zhang D, Guo G. Construction of a human cell landscape at single-cell level. Nature. 2020;581(7808):303–9. https://doi.org/10.1038/s41586-020-2157-4.
57. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The human cell atlas: from vision to reality. Nature. 2017;550(7677):451–3. https://doi.org/10.1038/550451a.
58. Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. J Mol Endocrinol. 2019;62(1):R21–45. https://doi.org/10.1530/jme-18-0055.

Chapter 4
Protein-Coding Genes

Luciana Amaral Haddad

4.1 Introduction

In the 1960’s, the laboratory of the 1968 Nobel laureate in Physiology or Medicine, Marshall W. Nirenberg, deciphered the genetic code by conducting *in vitro* polyribonucleotide chain (RNA) translation, owing to technological milestones, such as transfer RNA (tRNA) isolation and *in vitro* RNA synthesis, achieved by the 1968 Nobel co-laureates H. Gobind Khorana and Robert W. Holley [1]. This accomplishment meant the interpretation of the significance for protein synthesis of the 64 possible codons composed of triplets of the ribonucleotides adenine (A), cytosine (C), guanine (G) or uridine (U). It was hence demonstrated that 61 trinucleotide sequences code for amino acids, whereas the remaining three codons terminate protein synthesis (translation stop codons; Box 4.1). Considering that there are 20 possible amino acids found in natural proteins, it is logical that more than one codon may code for the same amino acid, a property known as the genetic code degeneracy. The single codon for methionine also signals the ribosome and tRNA where translation should start, thus recognized by far as the most frequently used translation start codon. As the genetic code rules are relatively simple and phylogenetically nearly universal (see Chap. 10), they easily lead to the computational prediction of the protein-coding sequence of genes or, simply, the coding sequence (CDS; Box 4.1). In such computational predictions, the output is the open reading frame (ORF; Box 4.1) for translation, a DNA sequence segment with a length in base pairs (bp) that is a number divisible by three. The ORF first and last nucleotide triplets must be the translation start codon and a stop codon, respectively, and all codons in

L. A. Haddad (✉)
Department of Genetics and Evolutionary Biology, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil
e-mail: haddadL@usp.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_4

Box 4.1 Key Concepts Related to Eukaryotic Protein-Coding Gene Sequence Elements

Term	Definition
Genetic code	A code for the correspondence between each of the 64 possible triplets of ribonucleotides and the 20 amino acids from natural proteins or termination translation signals.
Coding sequence (cds)	A sequence of DNA or RNA composed of nucleotides in multiples of 3, defined at 5′ by the translation initiation codon and at 3′ by one of the termination codons; in between there must be triplets of nucleotides corresponding to codons for amino acids. Translation of the CDS produces a polypeptide.
Open reading frame (ORF)	A gene analysis term employing the characteristics of the CDS for its identification on an unknown DNA sequence under study.
Exon	The gene sequence transcribed and maintained in the mature mRNA containing, in protein-coding genes, the CDS and UTRs. There are single-exon genes, and multi-exon genes in which exons are interrupted by intron sequences.
Intron	A gene sequence that intervenes between exons, is transcribed and removed from the pre-mRNA by splicing in the cell nucleus.
Transcription start site (TSS)	The first nucleotide (+1) transcribed by POLII defining the 5′ end of the gene and mRNA (to which the cap will be added).
Core promoter	A gene-associated element defined by the minimal sequence overlapping the TSS, extending nearly 40 nucleotides upstream and 40 nucleotides downstream, consisting of a variable combination of position-specific motifs necessary to activate transcription by POLII by associating with GTFs.
CpG islands	A gene-associated element with high C + G content, rich in CpG dinucleotides, associated with housekeeping genes and active chromatin. It is a frequent kind of promoter classified as dispersed, spanning in average 200–1000 nucleotides in total length, up- and downstream of the TSS, containing variable *cis*-acting elements that directly or indirectly interact with PIC.
Coding strand	The DNA chain of a protein-coding gene that contains the CDS.
Template strand	The DNA chain of a protein-coding gene that serves as template for transcription, and is complementary to the coding strand containing the CDS.
Antisense gene	A gene in the same chromosomal locus as a sense gene but with opposite orientation.
Protein-coding gene (eukaryotes)	The gene concept adopted here is the DNA sequence corresponding in length to its primary transcript, comprising the DNA strand that serves as template for transcription and the complementary chain, the coding strand, presenting the 5’-UTR, CDS that can be interrupted by introns, the 3’-UTR and polyadenylation signal.
Promoter-proximal element	A gene-associated DNA segment at −60 to −100 upstream of the TSS, containing *cis*-acting elements with regulatory roles on transcription initiation.

Term	Definition
Enhancer	A gene-associated *cis*-acting element with variable localization in regard to the TSS and action on either orientation leading to exponential activation of transcription of associated genes, by increasing the utilization of eukaryotic promoters.
Silencer	A gene-associated *cis*-acting element with variable localization in regard to the TSS and action on either orientation leading to repression of transcription of associated genes.
DNaseI hypersensitive site	A nucleosome linker DNA in open chromatin structure that is hypersensitive to digestion by DNase I, an endonuclease that cleaves DNA unbound to proteins.
5′-untranslated region (5’-UTR)	The exonic gene sequence or the sequence in the mRNA comprehending all nucleotides 5′ to the CDS.
3′-untranslated region (3’-UTR)	The exonic gene sequence or the sequence in the mRNA comprehending all nucleotides 3′ to the CDS (until the poly-A tail in eukaryotes).

between must be for amino acids. Although odds are high to computationally find a large number of ORFs on a given DNA sequence, a considerably long ORF generally corresponds to the CDS of a protein-coding gene, since over time that CDS has possibly undergone positive selection against base changes that introduce premature translation stop codons truncating the protein product. As one out of nearly 20 codons should be a translation stop codon, it is natural to assume that a long ORF should be indeed the CDS of a gene.

By the time when Nirenberg and colleagues cracked the genetic code, the human genome was believed to be largely composed of protein-coding genes, the total number of which was estimated to be considerably higher than it actually is. The draft of the human genome sequence presented in 2001 disclosed an estimate of nearly 30,000 total protein-coding genes [2]. The human genome project developed from 1989 in parallel with genome projects of model organisms. Upon sequencing of a new gene, identified features as presented in this chapter need to be annotated in association with its sequence deposited in standard databases that, as presented below, constitute a reliable and resourceful reference for genome analysis. It was then surprising to find out that the numbers of protein-coding genes did not differ significantly among multicellular organism genomes. Human gene annotation has been since 2001 continuously updated (see Sect. 4.3). Thus, the most recent assembly of the human genome sequence (hg38, annotation release 109 accessed in October 2020) displays 19,405 annotated protein-coding genes. In the data on multicellular animal genomes presented in Fig. 4.1, it is notable, for instance, that the fruit fly’s genome (0.16 Gigabase pairs, Gbp) is nearly 20 times smaller than the human genome (3.1 Gbp). The sizes of the illustrated genomes from other two invertebrates and six vertebrates show intermediate values. On the other hand, there

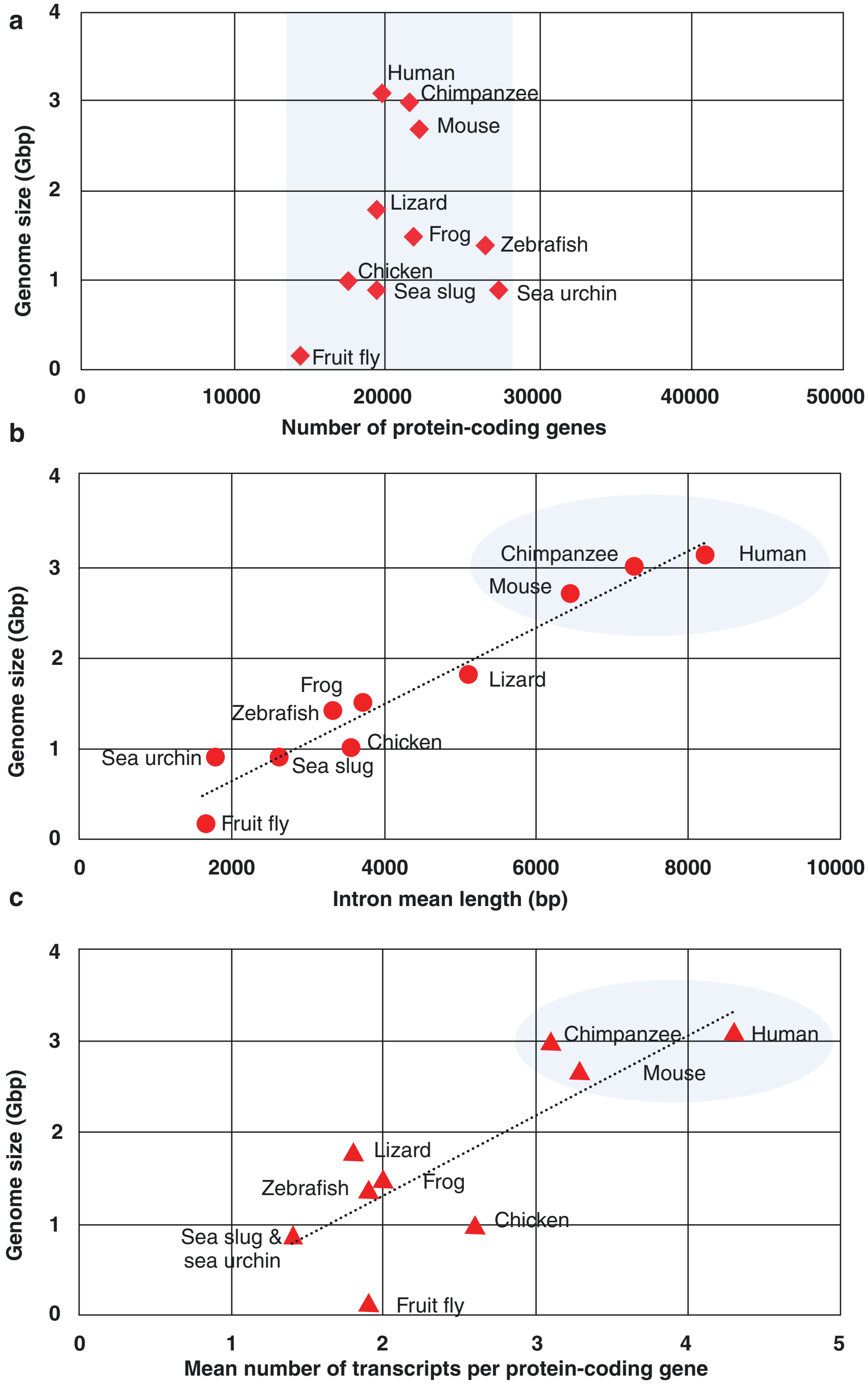

a
Genome size (Gbp)
0
1
2
3
4
Human
Chimpanzee
Mouse
Lizard
Frog
Zebrafish
Chicken
Sea slug
Sea urchin
Fruit fly
0
10000
20000
30000
40000
50000
Number of protein-coding genes
b
Genome size (Gbp)
0
1
2
3
4
Chimpanzee
Human
Mouse
Lizard
Frog
Zebrafish
Chicken
Sea urchin
Sea slug
Fruit fly
0
2000
4000
6000
8000
10000
Intron mean length (bp)
c
Genome size (Gbp)
0
1
2
3
4
Chimpanzee
Human
Mouse
Lizard
Zebrafish
Frog
Chicken
Sea slug &
sea urchin
Fruit fly
0
1
2
3
4
5
Mean number of transcripts per protein-coding gene

Fig. 4.1 (**a**) Graph relating the size of specific genomes (Gigabase pairs, Gbp) and the number of protein-coding genes. The grey color shadows the area where the numbers of protein-coding genes concentrate. (**b**) Graph relating the size of specific genomes (Gbp) and intron mean length (base pairs, bp). The grey color shadows the three mammalian species with the highest mean intron lengths. (**c**) Graph relating the size of specific genomes (Gbp) and mean number of transcripts per protein-coding gene. The grey color shadows the three mammalian species with the highest mean numbers of transcripts per protein-coding gene. Data were collected at the National Center for Biotechnological Information (NCBI, Bethesda, MD, USA; http://www.ncbi.nlm.nih.gov/genome) for the following species (common name, as indicated; genome assembly/annotation release): *Drosophlia pseudoobscura* (Fruit fly; UCI_Dpse_MV25/104), *Aplysia californica* (Sea slug; AplCal3.0/102), *Strongylocentrotus purpuratus* (Sea urchin; Spur_5.0/102), *Danio rerio* (Zebrafish; GRCz11/106), *Xenopus tropicalis* (Frog; UCB_Xtrop_10.0/104), *Anolis carolinensis* (Lizard; AnoCar2.0/102), *Gallus gallus* (Chicken; GRCg6a/104), *Mus musculus* (Mouse; GRCm39/ 109), *Pan troglodytes* (Chimpanzee; Clint_PTRv2/105), *Homo sapiens* (Human; GRCh38.p13/109). Accessed in October 2020

is only a nearly two-fold difference in the protein-coding gene amounts between the fruit fly (14,343) and the sea urchin (27,447) that represent the extremes in gene number on the graph (Fig. 4.1a).

Although the number of protein-coding genes does not correlate to the genome size among multicellular animals, the mean length of DNA sequences that intervene in the exon CDS, known as introns (see Sect. 4.2.4; Box 4.1), reports a direct relationship with the genome size (Fig. 4.1b), composing, in the case of the human gneome, nearly 25% of its sequence content (see Chap. 1). Together with gene introns, non-CDS repetitive DNA elements justify the increasing genome size along the Metazoan evolution, namely, interspersed repetitive DNA (see Chap. 8), copy number variations (see Chap. 9), simple tandemly repeated sequences as mini- and microsatellite DNA (see Chap. 6) or the heterochromatic DNA repeats that structurally compose the centromeres (see Chap. 6) and telomeres (see Chap. 7).

As there is no remarkable difference in the total of protein-coding genes among genomes from animal species in different classes (Fig. 4.1), this variable does not explain the differences in complexity for morphological and physiological phenotypes among invertebrate and vertebrate species. In fact, mechanisms that modify the gene expression likely account for the large diversity in animal phenotypes. Cell- and development-specific tuning of gene expression is highly modulated by molecular processes that are increasingly frequent along the phylogenetic scale, as diversity in *cis*-regulatory elements (see Sect. 4.2.1.2), alternative splicing (see Sect. 4.2.4.2) and RNA interference by microRNAs (see Chap. 5), which consequently lead to qualitative and quantitative variations in the full sets of RNA (transcriptome) and protein (proteome). One quantitative measure of alternative splicing is the mean number of transcripts per gene estimated from analyses of the species transcriptome that directly relates to the increasing phenotypic complexity in animal species (Fig. 4.1c).

4.2 Linear and Structural Elements Controlling Transcription and Primary Transcript Processing

In eukaryotes, effective transcription of protein-coding genes by RNA polymerase II (POLII) is characterized by consecutive initiation, elongation and termination phases co-occurring with the nuclear processing of the primary transcript, comprehending 5′ capping, splicing and 3′ polyadenylation. Hence, the processed transcript constitutes the mature 5′-capped and 3′-polyadenylated mRNA, consequently becoming ready to be exported to the cytoplasm where it may be translated (Figure 4.2a). As the primary transcript splicing takes place co-transcriptionally, the mature mRNA will be the output of both, the DNA transcription and the processing of the primary transcript, also known as pre-mRNA. Therefore, the combinatorial effect of different *cis*- and *trans*-acting factors will determine the production of the mRNA and its qualitative and quantitative aspects. For reviews on the eukaryotic gene expression steps and mechanisms, the reader may refer to different resources listed at the end of this chapter [3, 4].

4.2.1 Gene Transcription-Regulating cis *Elements*

In this section, we present the protein-coding gene *cis* elements involved in initiating and regulating transcription and, in section 4.2.4, we will discuss the *cis* elements defining exons, introns, and controlling the pre-mRNA processing. Transcription and pre-mRNA processing proceed by distinct large enzymatic complexes assembled by multiple molecular interactions. However, specific short DNA (for transcription) or pre-mRNA (for splicing and polyadenylation) sequences are central to each of those processes. We refer to these sequences as *cis*-acting elements as they belong to the reference molecule in the specific gene expression process and must physically interact with *trans*-acting protein or RNA factors to signal to the respective machineries of DNA transcription or pre-mRNA processing. Taking DNA transcription as reference, here we will discuss protein-coding gene DNA *cis* elements, which are the promoter and promoter-proximal elements, enhancers and silencers.

4.2.1.1 The POLII Promoter

There is no universal eukaryotic gene promoter. Although gene promoters have received different classifications, they may simply classify as focused, core promoter — ~80 nucleotides in length with a predominant transcription start site (TSS), frequently associated with genes with regulated transcription — and dispersed or diffuse promoters (multiple, dispersed TSS along 200–1,000 nucleotides in length, more often associated with genes with constitutive expression, the housekeeping genes) (Box 4.1).

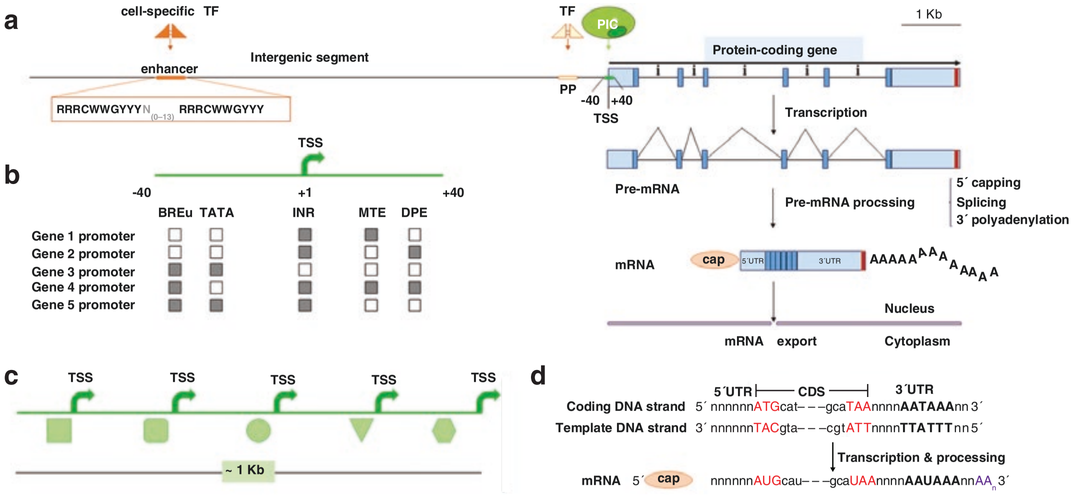


Fig. 4.2 A hypothetical eukaryotic protein-coding gene and its associated elements. (**a**) Diagram representing a gene with six exons (blue blocks) and five introns (i; lines between blue boxes) and an intergenic segment to the left (upstream) of the gene. In exons, the CDS is darker than the UTRs (5′ and 3′). The polyadenylation signal is indicated as a narrow red bar in the last exon. The length of the protein-coding gene is indicated by a long right-handed arrow, according to the length of its transcript and to the consensus in chromosome maps. The steps of transcription and pre-mRNA processing are sequentially presented. The gene-associated elements as the core promoter (−40 to +40), promoter-proximal elements (PP) and enhancer (similar representation to silencer) are shown according to their location in respect to the TSS. Their *trans*-acting factors are illustrated as PIC (for the green core promoter) or transcription factor (TF). The enhancer sequence is according to p53 response element consensus (R is purine, Y, pyrimidine, W is A or T). Scale bar for A: 1 kb. (**b**) A diagram of 80 bp (−40 to +40) of the core promoter is represented for five putative genes according to the presence among five motifs (BREu, TATA, INR, MTE and DPE). Their location is indicated according to the TSS. For this focused promoter, one predominant TSS is represented. Open and closed boxes mean absence and presence of the respective motif, respectively. (**c**) The 1-kb long CpG island illustrates a dispersed promoter containing five different motifs (distinct shapes), each one when activated will elicit transcription initiation in a different position, as indicated by right-handed arrows under each TSS. (**d**) Relationships among coding and template strands and mRNA. Red triplets of nucleotides represent the translation start and termination codons. The 5′ cap and 3′ poly-A tail are added to the mRNA post-transcriptionally; thus do not show equivalence in the DNA sequence. The polyadenylation signal remains in the mRNA. Abbreviation of the CDS, 5′- and 3’ UTRs are indicated by nucleotides (n)

The first transcribed nucleotide of a gene is the TSS or position +1 (Box 4.1). If all nucleotides 5′ to the TSS are numbered negatively in a descending sequence and those 3′ to the TSS numbered positively in an ascending sequence, one may consider the gene focused core promoter as the sequence lying nearly between the positions −40 and + 40. The core promoter position in relation to the TSS is mostly invariable, although there may be differences in its sequence span as it reflects the ensemble of *cis* elements represented within it (see below). The eukaryotic core promoter corresponds to a variable combination of sequence motifs (*cis* elements). There are nearly 15 DNA motifs known to occur at the −40-to-+40 core promoter sequence, conserved among eukaryotic genes. A common core promoter will have in average two (varying from one to five) of those motifs represented in that narrow sequence window. In spite of the large variability of promoter motif combinations, each individual motif shows a defined position in the −40-to-+40 window. The invariability of the location of the focused core promoter elements is due to the need of constitutively expressed proteins known as general transcriptions factors (GTF) to associate with them in order to place POLII appropriately to start transcription at the TSS. Thus, independently on the gene associated with a core promoter, the tridimensional volume of the macromolecular complex composed of the GTFs and POLII, known as the pre-initiation complex (PIC), will fit and associate with the DNA spanning the TSS (Fig. 4.2a and b). Thus, focused core promoter motifs are necessary to guarantee binding to at least one or two GTFs.

Each conserved motif of a core promoter has a consensus sequence determined due to the highest similarity of the base in a defined position upon alignment of multiple gene sequences covering the narrow window overlapping the TSS. For instance, TATA-box is a well-known core promoter element named upon the identification of the conserved consecutive bases T, A, T, and A at the core of the motif. TATA-box lies near the −31 to −26 position of the core promoter. It has long been demonstrated that the TATA-box *cis* element binds with high affinity to a *trans* factor, a protein (TATA-binding protein, TBP) that is one of the subunits of a GTF class (TFIID; transcription factor D of POLII). Although TATA-box is mostly a core promoter prototype in many textbooks, analyses after the Human Genome project and in the subsequent ENCODE project (see Sect. 4.2.5) disclosed that TATA-boxes are part of less than 40% of core promoters from human protein-coding genes, and that frequency is not very different from other Metazoan protein-coding gene promoters (Fig. 4.2b). The recent ‘post-genome’ analyses of core promoters revealed other conserved *cis* elements at the −40-to-+40 sequence as related to the TSS location. Some of them had been already recognized as transcription promoter elements, others were novel.

Out of almost two tens of known core promoter motifs, the one spanning the TSS (−2 to +5) known as initiator (INR) is apparently the most common *cis* element in focused promoters. It binds to TBP-associated factors (TAF) that as TBP are subunits of TFIID. There are DNA recognition motifs for TFIIB (TFIIB recognition element) that locate either upstream (BREu at −37 to −32) or downstream (BREd at −25 to −20) of the TATA box location. Moreover, there are core promoter motifs located downstream of the TSS, indicating they are transcribed and part of the 5′ untranslated region of the mRNA (see Sect. 4.2.4), such as the downstream core promoter element (DPE at +28 to +33) and the immediately upstream motif ten element (MTE at +18 to +29), both recognized by TFIID. MTE and DPE may depend on a functional INR to be efficiently recognized, and are enriched in TATA-less promoters. These are some of the most frequent core promoter motifs recognized by GTFs. As seen, there is no universal core promoter element and different compositions may be identified among genes [5] (Fig. 4.2b).

Dispersed (diffuse) promoters are better characterized as C plus G-rich sequences extending upstream and downstream of the TSS in average from 200 bp to 1 kb in total length (Fig. 4.2c), and displaying an observed frequency of cytosine-guanosine dinucleotides higher than expected in the adjacent bulk DNA. These promoters are widely known as CpG islands due to the high number of clusters of CpG dinucleotides, where ‘p’ stands for the phosphodiester bond between the cytosine and guanine. The dinucleotide CpG is statistically underrepresented in vertebrate DNAs, and the presence of CG-rich regions just upstream of TSSs is a distinctly nonrandom distribution.

The overall high frequency of non-methylated cytosines of CpG dinucleotides in CpG islands positively correlates with promoter activity (see Sect. 4.2.2). The CpG dinucleotide occurrence in clusters possibly creates binding sites for transcription factors [6]. DNA *cis* elements that commonly distribute within CpG islands include different focused core promoter motifs (see above), as well as motifs for transcription factors with ubiquitous expression, such as specificity protein 1 (SP1), E26 transformation-specific (ETS) protein and CREB-binding protein (CBP). SP1 recognizes GC-rich DNA elements based on the consensus sequence 5’ GGGGCGGGG 3′, and recruits the GTFs, specifically interacting with members of TFIID. ETS binds to the consensus sequence 5’ ACCGGA(A/T)GT 3′ and recruits co-activators as CBP connecting the transcription factor to the basal transcription machinery, assembling the transcriptional PIC. In summary, both focused core promoter and dispersed CpG island promoter represent a myriad of motif combinations that need to be specifically recognized by their *trans*-acting transcription factor to signal for PIC assembly and transcription initiation.

While the association of promoter motifs with PIC allows for TSS selection in active chromatin (see Sect. 4.2.2) and transcription initiation, the rate of RNA POLII initiation (escape from the promoter allowing positioning of a novel POLII unit on it) should be modulated by tissue-specific transcription factors. When these factors bind to their *cis* elements (see Sect. 4.2.1.2) and integrate with PIC through co-activators, for instance, the mediator and SAGA (Spt-Ada_Gcn5 acetyltransferase) complexes and CBP, the intensity of transcription initiation drastically modifies [7].

The activation of the protein-coding gene promoter permits POLII to initiate transcription of only one DNA strand, named the template (antisense or minus) strand (Box 4.1). The complementary strand, known as the coding (sense or plus) strand, is the DNA chain that contains the CDS with the correct successive sequence of codons from the translation start codon to termination codon (Box 4.1). Theoretically, when transcription of the DNA template strand sequence takes place, the synthesized single-stranded unprocessed pre-mRNA presents a sequence identical to that of the coding strand, except for the presence of uridine instead of thymine (Fig. 4.2d). However, as pre-mRNA processing occurs co-transcriptionally, 5′ capping is an early event before transcription elongation is engaged, and splicing each intron tends to proceed consecutively along transcription elongation (see Sect. 4.2.4.2).

On a chromosome map, an arrow should indicate the location of a protein-coding gene, starting at the promoter and pointing towards the end of the CDS according to the coding strand (Fig. 4.2a). In the same locus there may be another gene but with its coding strand on the opposite chain. Thus, one may consider these two genes as a pair of sense-antisense genes (see Chap. 5; Box 4.1). A locus containing differently overlapping sense and antisense genes may be referred to as a locus of nested genes. This terminology may be confusing as the coding strand can be referred to as the sense strand. For this reason, here we would rather define the DNA chain harboring the CDS simply as the coding strand, and apply the terms sense and antisense to genes in the same locus but with transcription in opposite directions.

Bidirectional transcription is the transcription initiated on two TSSs located less than one kilobase (Kb) apart, elongating in opposite directions on plus and minus strands of the DNA. Nearly 10% of the human genes are located in a head-to-head fashion as pairs in non-overlapping opposite directions, showing individual TSS located in the 1-kb bidirectional promoter. Bidirectional transcription of head-to head gene pairs is frequently tissue-regulated in a coordinated manner (see Sect. 4.2.3). These bidirectional promoters appear symmetrically enriched in CpG islands and BRE motifs. By contrast, TATA motifs are enriched in unidirectional promoters and, in bidirectional promoters they are asymmetrically distributed near one TSS. INR and DPE motifs distribute equally between uni- and bidirectional promoters [7]. Because promoter sequences can overlap the regulatory sequences of other genes, we here adopt a generally employed concept that promoters are gene-associated sequences and not part of the gene, which are considered here as the DNA sequence corresponding to the total length of the primary transcript (Box 4.1).

By contrast, it has been unveiled that nearly 50% of the promoters in the human genome can activate transcription bi-directionally, although in most cases a protein-coding gene is found in only one direction. In these cases, short noncoding RNAs (ncRNAs) spanning either the promoter or upstream are produced in the direction that lacks a gene. Although these ncRNAs appear as markers of active promoters, it is yet unclear if they are functional, as they have short half-lives, being thus barely detectable and likely prone to degradation (see Chap. 5) [7].

4.2.1.2 Transcription Regulatory Elements

Gene transcription initiation activated at the core promoter allows for basal levels of pre-mRNA synthesis. However, cell identity depends on a tightly controlled transcription program permitting increased expression of a subset of genes while others remain silent or with very low to low expression levels.

Trans-acting transcription factors specifically expressed in a cell type regulate transcription of protein-coding genes by interacting with gene-associated DNA *cis* elements in the promoter-proximal region or the variably located enhancers and silencers. For a drastic change in gene transcription initiation at the TSS upon activation of the promoter, the cell-specific transcription factors bound to *cis*-regulatory DNA elements need to interact with PIC, a communication allowed by the large mediator protein complex, co-activators or co-repressors, often eliciting chromatin looping [4].

The gene expression regulatory *cis* elements located close to the core promoter at −60 to −100, collectively known as promoter-proximal elements (Box 4.1 and Fig. 4.2a), include consensus sequences as the CCAAT box, the GC-box (5’ GGGGCGGGG 3′) recognized by SP1 (see Sect. 4.2.1.1), and the octamer (5’ AGCTAAAT 3′) element, among other *cis* elements. The CCAAT box is an important regulatory element at −60 to −100 upstream of the TSS and the core promoter. The heterotrimeric NF-Y protein is an established CCAAT-binding protein controlling gene expression mostly independently on the cell type [8] (see Sect. 4.2.2 and Sect. 4.3.3).

OCT proteins, members of the POU (Pituitary-specific factor, Octamer-binding factors OCT1 and OCT2 and UNC-86) domain-containing family of transcription factors, recognize the octamer element, and are important to establish and maintain cell fate and cell identity throughout the embryonic development. The protein OCT4, for instance, is expressed in mouse and human totipotent germ cells and pluripotent embryonic stem cells. Its major roles in pluripotency maintenance have been demonstrated by different researchers, including the group of one of the 2012 Nobel laureates in Physiology or Medicine, Shinya Yamanaka, who demonstrated that the overexpression of the *Oct4*, *Sox2*, *Myc* and *Klf4* genes in adult mouse or human fibroblasts induces the characteristics of embryonic stem cells by reprogramming pluripotency, a protocol now widely used to establish *in vitro* induced pluripotent stem cells (iPSC) [9]. Evidence suggests that OCT proteins may activate transcription of target genes by directly interacting with TBP and recruiting the chromatin remodeling complex (see Sect. 4.2.2).

Additional transcriptional factors such as those of the large homeotic factor family HOX may also specifically recognize their cognate *cis* elements in the promoter-proximal region immediately upstream of the TSS and the core promoter. Hence, different promoter-proximal transcription regulatory elements associate with transcription factors that activate expression of protein-coding genes of utmost importance for the embryo development. Although some promoter-proximal elements may be recognized by ubiquitous transcription factors, here we opt to classify this region as a proximal segment of regulatory elements because, differently from the core promoter and CpG islands, many motifs within it are activated by tissue-specific transcription factors exponentially modifying the levels of transcription.

Transcription regulation may be achieved by integration of more distant *cis* elements, enhancers or silencers, bound to tissue-specific protein factors acting as activators or repressors of transcription, respectively (Fig. 4.2a). Enhancers activate transcription in levels considerably higher than basal transcription (Box 4.1) [7]. Conversely, silencers recognition by transcription factors repress transcription possibly leading to chromatin remodeling of the promoter of the target genes (Box 4.1) [10]. *Cis*-regulatory motifs often consist of two short direct or inverted DNA repeats, separated by a limited central sequence, consistent with binding by transcription factor dimers (Fig. 4.2a and Sect. 4.2.3). Chromosome folding puts these farther elements close to PIC in a three-dimensional space also occupied by larger complexes such as the mediator and the chromatin remodeling complex (see Sect. 4.2.2). One characteristic of these regulatory elements is that they may be upstream or downstream of the TSS, being inter- or intragenic, not uncommonly within introns. As they may reside within a gene but control transcription of up- or downstream gene(s), enhancers and silencers are generally classified as gene-associated elements.

The distance of enhancers to the TSS increases with genome growth in size and complexity. Thus, the majority of enhancers lie within 10 Kb from the promoter of genes of *Drosophila melanogaster* while, in the human genome, distances as long as 1 Mb have been reported and often spanning few hundreds kilobases, although 50 kb has been the considered average. Moreover, enhancers have been abundantly identified in the human genome accounting for estimations of nearly hundreds of thousands units. By contrast, silencers have been understudied and few thousand units have been characterized so far [10]. It is thus perceived that the increase in genome size contributes to diversify the *cis*-regulatory elements impacting gene transcription probably by accommodating longer chromatin loops, consequently playing roles in complex phenotypes. In addition, although the estimates for enhancer-TSS associations based on chromatin analysis vary according to tissue and developmental phase studied, some reports disclose approximately four enhancers acting upon a single gene in the human genome, and in average 2.5 TSSs selected by a given distal *cis*-regulatory element, except for clusters of developmentally regulated genes (see Sect. 4.3.3). When chromatin analyses are associated with expression studies across cDNA libraries, enhancers appear remarkably redundant in respect to transcription patterns but synergistic regarding expression strength [7].

Although there has been a functional enhancer/promoter dichotomy in the literature, different studies have reported that enhancers may directly recruit PIC to initiate transcription, producing small amounts of RNA (eRNA; enhancer-associated RNA) that in general are capped, short, unspliced and in a smaller scale than that initiated at promoters. These molecules, eRNAs, resemble the antisense transcripts from bidirectional promoters containing a single sense protein-coding gene (see Sect. 4.2.1.1). In addition, eRNAs are mostly nuclear, non-polyadenylated, and susceptible do degradation yet considered as markers of enhancer activity. Enhancers that produce eRNAs tend not to overlap exons of known genes, indicating they should not be coincident with alternative promoters (see Sect. 4.2.3). The production of eRNAs appears to maintain open chromatin in the enhancer locus (see Chap. 5) [7].

4.2.2 Chromatin Remodeling

The synergistic functional interactions among DNA cytosine chemical alterations, histone variant expression and post-translational modifications, the transcriptional machinery, tissue-specific transcription factors, and chromatin remodelers play determining roles at the transcription level. The *cis* elements that initiate and regulate transcription are embedded in a chromatin structure fundamental in modulating their activity and accessibility to cognate *trans* factors.

Active promoters are associated with specific chromatin signatures, as nucleosome-depleted regions (NDR; reduced nucleosome occupancy over promoters) directly defining DNaseI hypersensitive sites (DHS) in the nucleosome linker DNA (Box 4.1 and Fig. 4.3a), and enrichment of specific histone isoforms and post-translational modifications. Accordingly, an additional way to classify gene promoters (see Sect. 4.2.1.1) is based on specific histone post-translational modifications and DHS presence controlling chromatin accessibility to transcription factors.

Nucleosomes correspond to the first level of chromatin compaction (see Chap. 2), and are its essential functional unit (Fig. 4.3a). As nucleosomes are thermodynamically stable structures, active remodeling enzymes and passive non-histone proteins are needed to reposition nucleosomes. Nucleosome remodelers are ATP-dependent enzymes that use the energy of ATP hydrolysis to disrupt crucial DNA–histone interactions (Box 4.2). They basically group into two main classes, the nucleosome translocation enzymes that slide histone octamers along DNA, and histone exchange factors, which physically remove the entire histone core or exchange histone sub-complexes for a histone variant.

Box 4.2 Key Epigenetics and 'Post-Transcriptional' Processing Concepts Related to the Expression of Eukaryotic Protein-Coding Genes

Term	Definition
Chromatin remodelers	Two major classes of ATP-dependent enzymes (nucleosome translocation enzymes and histone exchange factors) that actively modify nucleosome positioning.
Nucleosome positioning	Localization of nucleosomes in regard to gene-associated *cis* elements dependent on epigenetic marks and transcription activity status.
Cytosine methylation	Methylation of cytosine by DNA methyltransferases consequently eliciting chromatin condensation by recruitment of chromatin modifiers.
Histone tail posttranslational modifications	Covalent modifications that are added after translation to specific amino acid side chains (eg. Lysine –K) of histones by specialized enzymes.
Insulator	A chromatin boundary element or barrier that can block the spreading of heterochromatin from an adjacent region and presents enhancer-blocking activity.
Enhancer-blocking activity	A transcriptional *cis* regulatory region that when located between an enhancer and a gene's promoter prevents the enhancer from modulating the expression of the gene.
Locus control region	A DNA region that includes DNase hypersensitive sites 5′ to a gene or gene cluster that confers high-level, position-independent, and copy number-dependent expression of that gene or gene cluster.
5′ capping	One of the three steps of the primary transcript processing consisting of the addition of a guanosine to the 5′ end of eukaryotic mRNAs and its methylation, being recognized as the 5' cap.
Constitutive splicing	One of the three steps of the primary transcript processing consisting of removal of all intronic sequences and consecutive ligation of all exons.
Alternative transcript	A mature mRNA sequence that differs from the reference due to alternative splicing (intron retention, altered exon composition affecting the CDS and or the UTRs, or RNA circularization) or alternative promoter.
Protein isoform	Amino acid sequence of a protein chain that differs from the expected translation of the respective reference gene CDS due to alternative splicing, alternative translation initiation, regulated proteolysis or expression by a closely related paralogous gene.
3′ polyadenylation	One of the three steps of the primary transcript processing consisting of recognition of the polyadenylation signal in the nascent pre-mRNA, its cleavage nearly 20 nucleotides downstream and activity of the poly(A) polymerase synthesizing a poly-A tail in the 3' end of the mRNA.
Gene imprinting	Expression of a single allele of a diploid gene locus on a parent-of-origin-dependent basis due to regulated epigenetic silencing of the other allele.
Imprinting control region	A regulatory region that controls epigenetic imprinting and affects the expression of target genes in an allele- or parent-of-origin-specific manner.

The density and exact positioning of nucleosomes correlate with transcriptional regulation. While the CDS is typically dense in nucleosomes, in active promoters accessible TSS in NDR is flanked by two well-positioned nucleosomes, the upstream −1 nucleosome and the downstream +1 nucleosome (Fig. 4.3c; Box 4.2). Recent studies suggested that the transcription factor NF-Y bound to the CCAAT box in the promoter-proximal segment controls the fidelity of transcription initiation at gene promoters in mouse embryonic stem cells by maintaining the region upstream of the TSS in the NDR while simultaneously protecting this accessible region against ectopic transcription initiation [8]. In addition, the +1 nucleosome forms a barrier, beyond which nucleosomes are packed, resulting in uniform positioning, which decays at distances farther from the barrier.

4.2.2.1 Cytosine Methylation

The paucity of regions rich in CpG dinucleotides in mammalian genomes outside CpG islands strongly contributes to nucleosome positioning, as CpG islands generally compose NDRs. One key factor that may justify the scarcity of CpG dinucleotides in the rest of the genome is that methylated cytosines tend to undergo spontaneous deamination leading to thymine. It also explains the higher rates of genomic DNA substitutions of CpG cytosine over other dinucleotides. The differentially high affinity of specific proteins to methylated cytosines of CpG dinucleotides in CpG islands (see below) should avoid spontaneous deamination of these cytosines. As CpG islands are generally unmethylated, they preserve their high CpG dinucleotide content, and methylation of CpG island cytosine by DNA methyltransferases (leading to 5′-methylcytosine and 5′-hydroxy-methylcytosine) associate with transcriptional repression and is thus considered an epigenetic event [6].

Binding of proteins, such as methyl-CpG-binding protein 2 (MeCP2), to methylcytosine integrates its DNA sequence into protein complexes that include histone modification enzymes (eg., histone deacetylase complex) leading to dynamic changes in chromatin structure increasing compaction that impedes access of transcription factors (Box 4.2) [6]. Methylation of cytosine is responsible for the inactivation of one X-chromosome in cells of the mammalian female producing dosage compensation and the heterochromatic Barr body (see Chap. 2), silencing the expression of transposons and other interspersed genomic repeats (see Chap. 8), and repressing CpG island promoters due to genomic imprinting (see Sect. 4.5), dynamic mutations as in fragile X syndrome (see Chap. 6) or blocking the expression of tumor suppressor genes in cancer.

4.2.2.2 Histone isoforms and Post-Translational Modifications

The nucleosome consists of a histone octamer, composed of two units of each of the four histones H2A, H2B, H3, and H4 wrapped by nearly 147 bp of DNA, constituting a fiber of 11 nm in diameter (Fig. 4.3a). It is the binding of a fifth member of the

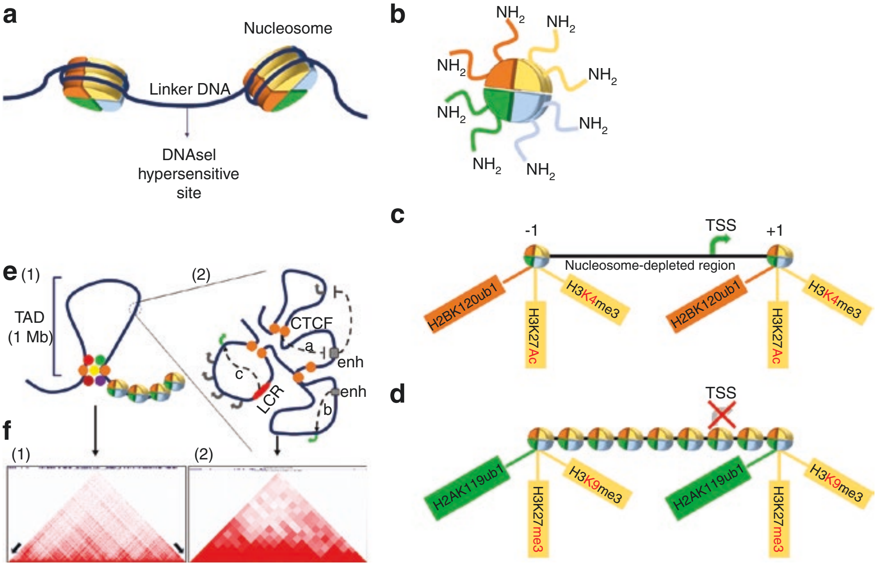


Fig. 4.3 (**a**) Drawing of two nucleosomes with a linker DNA representing the nucleosome-free DNA susceptible to DNAseI digestion. (**b**) A magnified drawing of the nucleosome octamer using four colors to represent the four histones, and their N-termini protruding away from the octamer. (**c**) Representation of an active promoter in NDR showing activity of the TSS, between the −1 and + 1 nucleosomes, and illustration of octamer histone exemplified by three modifications related to transcriptional activity. (**d**) Representation of a repressed promoter with condensed chromatin illustrated by histone octamers and histone modifications related to transcriptional repression. (**e**) (1) A TAD drawing depicts its insulation from heterochromatic region (arrayed nucleosomes on the right) by multiple proteins aligning two insulators defined by the CTCF (orange circles). (2) A short segment of the TAD from (1) is illustrated as TAD subdomains showing CTCF bound to insulators that although does not block chromatin show enhancer-blocking activity (a). An additional loop has an enhancer active over one promoter (b) that could be a small-scale (kb) region with chromatin contacts as illustrated in F2. In another insulated TAD subdomain, LCR controls the activation of five clustered genes (indicated by arrows as their promoters) with one (green promoter) activated at a time (c). (**f**) Heat maps (triangle) of chromatin interaction analysis of a human embryonic stem cell line, related to the tridimensional proximity pairs of loci that are adjacent in tridimensional genomic coordinates represented on both triangle adjacent sides. In (1), 1-Mb segment of human chromosome 17p13.1 containing nearly 65 genes is illustrative of TAD and two thick arrows indicate insulator location. In (2) the chromatin contacts detected in the 20-kb segment of the *TP53* gene are illustrative of TAD subdomain. Data obtained at the University of California in Santa Cruz (UCSC) genome browser accessed in November, 2020

histone family, histone H1 that allows the 11-nm fiber to tighten and coil the segment of nucleosomes increasing its diameter to 30 nm (see Chap. 2). Each octamer core histone has a structured core domain that binds DNA and a disordered N-terminal tail that projects into solution (Fig. 4.3b). A subset of histone-encoding genes expresses histone variants believed to replace its paralog in the nucleosome octamer under certain physiological circumstances. For instance, histone variant H2A.Z occupancy at CpG islands increases upon transcriptional activity suggesting that its deposition in nucleosomes flanking the TSS should control transcription and be attributed to low CpG island methylation levels [11].

Differential chromatin folding is primarily elicited by specific DNA-histone interactions modulated by histone posttranslational modifications. Besides the ATP-dependent chromatin remodeling complexes, the epigenetic machinery contains enzymes that post-translationally modify the side chain of specific histone residues, as histone acetyltransferases (HATs), histone deacetylases (HDACs), histone kinases and methyltransferases, among others. Different histone tail post-translational chemical modifications have been reported. However, acetylation, methylation, ubiquitination, and sumoylation have been more studied, and described to most often affect chromatin condensation. Histone modifications function as docking sites for chromatin readers that specifically recognize these modifications and recruit additional chromatin modifiers and remodelers to directly affect the chromatin condensation and gene transcription status (Box 4.2) [11, 12].

Nucleosome positioning and histone modification states can be used to classify promoters associated with different types of transcription initiation patterns. A vast variety of chemically modified residues have been described in the N-termini of the four histones that define the nucleosome octamer. Furthermore, modification of residues in the C-terminus is also commonly described in addition to some post-translational modifications in amino acids of the core of histones. As a general rule, however, active genes typically carry high levels of histone N-terminal lysine di- and tri-methylation, and acetylation on H3 and H4 N-terminal tails. At the downstream edges of promoter-associated NDRs, histone H3 tri-methylated at lysine (K) four (H3K4me3) within well-positioned +1 nucleosomes highly coupled with CpG islands has been shown to stimulate PIC formation. Studies suggest that ubiquitylation of H2B lysine 120 (H2BK120u1) stimulates H3K4me3, which promotes downstream H3/H4 acetylation (eg. H3K27Ac) through recruitment of HATs. H3K27me3 is associated with gene repression, while H3K27ac is associated with gene activation and active enhancers. Since they act on the same lysine residue, these marks are mutually exclusive. Repressed promoters associate, for instance, with H3K27me3, trimethylation of H3 lysine 9 (H3K9me3), and ubiquitylation of H2A on lysine 119 (H2AK119ub1) (Fig. 4.3d). By contrast, poised chromatin domains bear both the activation-associated histone modification H3K4me3 and the repression-associated modification H3K27me3, as in CpG islands linked to developmentally regulated genes in embryonic stem cells. Moreover, H3K27ac and H3K4me1 associate with active enhancers, and the bodies (exons and introns) of active genes are enriched in H3 and H4 acetylation, H3K79me3, and H2BK120u1, and increasing amount of H3K36me3 toward the 3′ end [11, 12].

4.2.2.3 Insulators, Locus Control Regions and Enhancer Blocking

Mammalian interphase chromosomes are highly heterogeneous, organized into discrete nuclear territories and globular domains of active and inactive chromatin. Euchromatin and heterochromatin exhibit intra- and inter-chromosomal contacts most frequently between segments with the same chromatin configuration, as well as attachment points of the heterochromatin with the nuclear lamina. In a narrower range, chromosomes are partitioned into near 1-Mb segments that tend to self-associate and insulate relatively from neighboring domains forming topologically associating domains (TADs; Fig. 4.3e1). Consequently, sequences within TADs tend to interact with each other more frequently than they do with sequences throughout the rest of the genome, and *cis*-regulatory elements as enhancers and silencers will commonly control transcription and chromatin of associated genes residing in the same TAD. *Cis*-regulatory elements often regulate not only a single gene, but a group of genes within a TAD. The conservation of TADs among different cell lines and across species suggests they define a layer of chromosome organization and folding. TADs are insulated by their borders consisting of the broadly expressed zinc finger nucleic acid-binding protein CCCTC-binding factor (CTCF) bound to its *cis* element (Fig. 4.3e1). Deletion of this element is sufficient to lead to loss of physical insulation and subsequent integration of two adjacent TADs into a single domain. A multimeric protein complex may assemble onto CTCF-containing borders of the genomic segments that limit the extent of TADs, constituting the insulators. Aligned insulators bound to the multimeric insulator protein complex preferentially interact with one another. TAD borders correlate with protein occupancy, relatively few chromatin interactions, high gene density, and active transcription [6, 12].

Chromatin insulators contribute to nuclear and genome organization as a barrier that can block the spreading of heterochromatin from an adjacent region and present enhancer-blocking activity (Box 4.2). Conversely, other associations between CCCTC *cis* element and CTCF may display enhancer-blocking activity but no barrier function (Box 4.2 and Fig. 4.3e2). It is the case of TAD subdomains (subTADs or contact domains), which are smaller regions within TAD's organization presenting enriched chromatin interactions occurring over shorter genomic distances at the gene (kb) scale. These subdomains direct specific gene-regulatory outcomes, either by facilitating or disrupting enhancer–promoter communication. Hence, consistent with additional compartmentalization within TADs, a single TAD can have both repressive and active chromatin signatures. Also, as sub-TADs vary along cell differentiation, they may represent cell-type specific long range enhancer-promoter contacts (Fig. 4.3e2).

Locus control regions (LCR) organize the expression of an entire gene cluster into an active chromatin domain developmentally enhancing the transcription of individual genes (Box 4.2). LCRs have been described in clusters of genes of the human alpha-globin (16p13.3), beta-globin (11p15.4), opsin light-absorbing visual pigment (Xq28) and the growth hormone 1 (17q23.3) families, among others. The segment containing an LCR and the gene cluster it controls is flanked by CTCF-binding sites (Fig. 4.3e2). A pair of CTCF sites will only engage in contact above local background, insulating the region, if they are in a convergent linear orientation (see Sect. 4.3.3). Change of orientation may lead the looping to redirect and disrupt packaging of chromatin [13].

The study of LCR and other TAD subdomain long-range looping contacts across multiple spatial scales is informative on the regulation of gene expression. Those contacts may be identified by high throughput DNA sequencing following affinity purification of covalently ligated DNA-protein cross-links isolated upon cell fixation, a technique known as Hi-C. Solid bioinformatics analysis of Hi-C data should yield an output of mapped sequences that have been crosslinked due to tridimensional proximity. The HiC layout may be presented as heat maps displaying a strong diagonal that shows the proximity pairs of loci that are adjacent in tridimensional genomic coordinates on adjacent sides as, for instance, in a triangle layout (Fig. 4.3f).

4.2.3 The **TP53** *Gene: A Genetic Case Study on Promoters and Enhancers*

The tumor suppressor protein 53 (p53), encoded by the *TP53* gene (17p13.1, OMIM 191170), is a transcription factor that controls cell cycle progression, DNA repair, target gene transcription, and genome stabilization, consequently modulating biological processes such as apoptosis, autophagy, and cell senescence. This 53-kDa protein has six domains: two transactivation domains, a proline-rich domain, a central DNA-binding domain, and regulatory and oligomerization domains at the C-terminus. It is thus an appropriate example to illustrate the theme of gene-associated elements regulating transcription.

4.2.3.1 *Trans*-Acting p53 effects on Gene Transcription

As a transcription factor, p53 dimer enters the nucleus, where it subsequently associates with another dimer resulting in a tetramer, before binding to its response element in target genes. *Cis* elements recognized by p53 may act as enhancers or silencers depending on other interacting proteins, and most often are upstream of the TSS, some within the promoter-proximal region, although p53 binding sites can be downstream of the TSS as well. It is believed that elements in the promoter-proximal region bound to p53 associate more frequently with activation of the downstream gene. It is also considered that when p53 action on the gene transcription is repressive it should be an indirect effect not associated with its *cis*-regulatory elements, but interacting with additional effective transcription factors. P53 indirect regulatory mechanism most commonly involves its direct target *CDKN1A* that encodes p21, which leads to reactivation of the cell cycle repressor complexes as the one mediated by the tumor suppressor retinoblastoma protein.

The canonic p53-responsive element comprises two similar decamers (5’ RRRCWWGYYY 3′, where R is purine, Y is pyrimidine, and W is A or T), separated by none to 13 variable nucleotides (Fig. 4.2a). In addition to its direct or indirect effects on promoter activation, p53 has been assigned putative roles in assisting and pioneering chromatin activity. P53 is known to activate transcription of genes for pro-apoptotic proteins and to repress transcription of genes encoding proteins that lead the progression of the cell cycle. Thus, when genotoxic stresses accumulate, there is an increase in transcription of the *TP53* gene and translation of its mRNA, and synthesized p53 regulates transcription of a large gene network that ultimately determines if the cell cycle arrests at late G1/early S phase for DNA repair or if an apoptotic pathway is activated leading to cell death [14].

4.2.3.2 *TP53* Gene Transcription

Many *trans*-acting factors have been identified to activate transcription of the *TP53* gene under genotoxic stress magnifying the amount of mRNA produced by binding to specific elements in this gene's promoter-proximal region. Recent observations have disclosed that *TP53* promoter lies less than 1 kb upstream of the neighboring gene *WRAP53* (*WD repeat-containing antisense to TP53*) promoter. Thus, *TP53* and *WRAP53* genes are positioned in a head-to-head fashion and share a bidirectional promoter. In fact, *TP53* and *WRAP53* first exons partially overlap (Fig. 4.4). There are no CAAT or TATA motifs upstream of *TP53*, but an INR consensus as well as GC-rich content and one SP1-binding site. Two CpG islands have been identified by the ENCODE project (see Fig. 4.4 and Sect. 4.2.5). Transcription initiation of these two genes appears to employ independent motifs, and distinct mapping of the 5′ end of each gene's transcripts disclosed a sharp predominant TSS for *TP53* and at least five different TSSs for *WRAP53* (Fig. 4.4, red arrows on the map below the bidirectional promoter). *WRAP53* transcripts appear to stabilize *TP53* transcript 5′ end by antisense hybridization. Chromatin activity signatures seen in Fig. 4.4 (DNAse hypersensitive sites, H3K4Me1, H3K4Me3 and H3K27Ac) immediately upstream of *TP53* are subject to regulation according to DNA damaging sensing and p53 activity.

As shown in Fig. 4.4, the 5′ ends of the shorter dark blue *TP53* transcripts do not align with the 5′ ends of the long transcripts initiated at the bidirectional promoter. The synthesis of those downstream *TP53* transcripts is due to alternative promoters in introns 1 and 4. Specifically in intron 4, promoter and enhancer motifs are conserved among species, inferring regulatory roles. Indeed, one of the motifs conserved in intron 4 is the p53-response element itself indicating an auto-regulatory loop.

P53 isoforms produced by mRNA transcribed from alternative *TP53* promoters are truncated at the N-terminus, deleting 133 (Δ133p53 isoform) or 160 (Δ160p53 isoform) amino acids. As seen, the N-terminus of p53 harbors the transactivation domain, absent in Δ133p53 and Δ160p53 isoforms, which may tetramerize with full-length p53 isoforms, compromising the transactivation of target genes (see Sect. 4.2.3.3). Expression of Δ133p53 and Δ160p53 isoforms has been reported in many malignancies and upon demethylation of CpG dinucleotides within *TP53* alternative promoters induced by DNA damage. Thus, besides *TP53* pathogenic DNA variants (see below) p53 activity may be modulated *in vivo* due to cell environmental changes leading to alternative expression and tetramer formation with different isoforms [15].

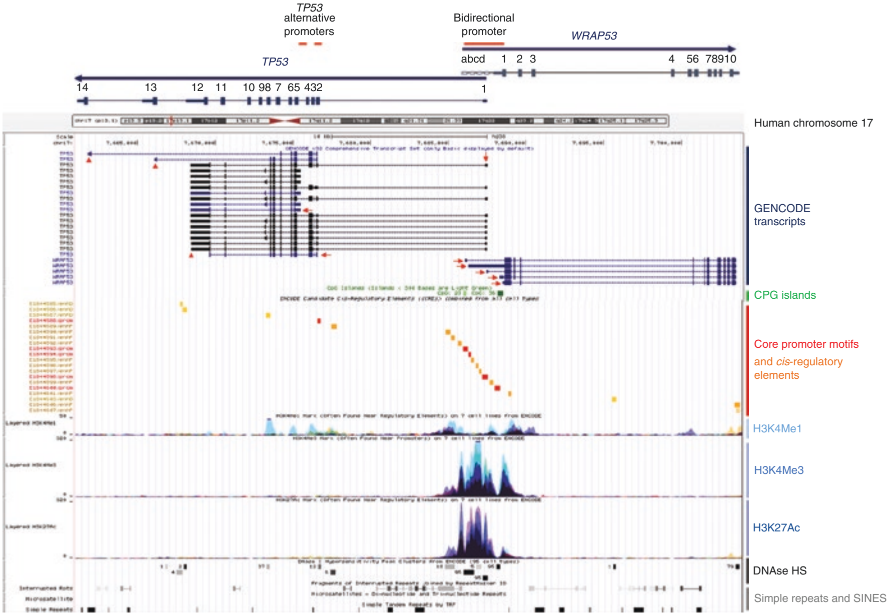


Fig. 4.4 A diagram view of a small part of human chromosome 17 band p13.1 (17p13.1) indicated by a red bar on the chromosome drawing, spanning 43 kb (7,661,001 to 7,704,000). The map displays some of the ENCODE project features for the mapped genes *TP53* and *WRAP53*, which are oriented in a head-to-head fashion, respectively antisense and sense. Alternative transcripts, verified by GENCODE V32 Comprehensive Transcript Set, are displayed beside 'GENCODE transcripts' for each of the two genes, having exons and introns indicated by boxes and lines, respectively. The part of the exons containing the CDS is higher than exon UTRs. As the *TP53* gene has at least two alternative promoters (indicated near exon 2 and in intron 4), transcription initiated at it produces the blue transcripts aligned with this gene among other short black transcripts. The long black transcripts had transcription initiated at the upstream promoter. *WRAP53* transcripts aligned with this gene are in blue. The 5′ ends of *TP53* and *WRAP53* genes overlap in their bidirectional promoter (indicated), in which there are two CpG islands indicated in green showing their numbers of CpG dinucleotides. Light green CpG islands are shorter than 300 bp. *WRAP53* transcripts show different TSS (red arrows) characteristic of dispersed promoter, whereas *TP53* transcripts show a predominant TSS indicative for core promoter. Mapped promoter and *cis*-regulatory element motifs (transcription enhancers in all cases in this figure) are presented, respectively, as red and orange boxes. Marks of H3K4Me1, H3K4Me3, H3K27Ac shown as frequency peaks, obtained from seven ENCODE-standardized cell lines, are enriched in the region of the bidirectional promoter, as indicated. DNAse hypersensitive sites (DNAse HS) are indicated in grey to black intensity proportional to the number of cell lines (out of 95) in which they had been identified at that position. Three red arrowheads indicate on *TP53* transcripts the localization of three polyadenylation signals in distinct exons. The location of simple repeats and interrupted short interspersed repeats are presented, and all of them map to introns or exon UTRs. Data obtained at the University of California in Santa Cruz (UCSC) genome browser accessed in November, 2020

4.2.3.3 Germline and Somatic *TP53* Pathogenic DNA Variants

TP53 is the human gene most frequently mutated in sporadic cancers, and germline pathogenic DNA variants in *TP53* cause Li-Fraumeni syndrome, a hereditary condition characterized by predisposition to multiple cancers. The Knudson's hypothesis [16], based on studies on retinoblastoma patient families, predicted that bi-allelic inactivation of a tumor suppressor gene is necessary for tumor development. Earlier onset of tumor development is observed for patients with inherited cancer predisposition syndromes. These patients inherit from one parent one inactivated allele, which is the germline pathogenic variant that will be present in all somatic cells of the individual. Thus, for the generation of a tumor cell clone the wild-type allele must undergo the second mutational hit in somatic cells leading to complete loss of function of the tumor suppressor gene. The need for only one new somatic loss-of-function DNA variant in hereditary cancer anticipates the occurrence of neoplasia in those patients. Moreover, the germline pathogenic variant in cancer predisposition syndrome patients increases penetrance over the lifetime of the individual, when compared to sporadic cancers due to two *de novo* somatic mutational events in a tumor suppressor gene in the same cell.

P53 regulates transcription by its two specialized transactivation domains interacting with the transcriptional machinery. It recognizes its DNA response *cis* elements by an elaborate mechanism involving a sequence-specific DNA-binding domain and the regulatory C-terminal domain. P53 tetramerization by its oligomerization domain is essential for stabilization of the p53–DNA complex, allowing each dimer to bind to one decameric sequence of p53-responsive *cis* elements associated with the target gene. The vast majority (~75%) of germline and somatic *TP53* pathogenic DNA variants are missense substitutions in exons 4 to 9, which encode p53 DNA-binding domain. These variants do not render p53 prone to degradation, but affect its ability to bind DNA, maintaining its property to oligomerize with wild-type p53. Consequently, mutant p53 successfully tetramerizes in a dominant-negative manner with its wild-type counterpart, preventing wild-type p53 functions. Accordingly, wild-type p53 functions are antagonized when included in tetramers containing mutant p53, which potentially gain new functions in preneoplastic lesions acting in a dominant negative way. The new functions of tetramers of wild-type and missense p53, though not fully understood, might be due to changing the binding affinity to DNA or associating with other proteins as transcription factors, switching transcription output of the gene network [14]. In the multistep process of malignant transformation, the *TP53* heterozygous state of cells is transient as mutation of the wild-type *TP53* allele may follow, leading to tumor initiation likely due to attenuation of the wild-type allele function below a critical threshold.

4.2.4 *Gene Elements Driving the Primary Transcript Processing*

Once transcription initiates and the pre-mRNA emerges from the RNA POLII, capping of its 5′ ribonucleotide takes place (Fig. 4.2a), consisting of an N7-methylguanosine linked to the 5′ end of mRNA by an unusual 5′-5′ triphosphate bond (Box 4.2). Translation of eukaryotic mRNA is mostly dependent on the 5′ capping for initiation. 5′ capping is also important to protect the 5′ end of mRNA against the action of 5′-to-3′ exonucleases allowing the mature mRNA to adopt a closed configuration by interaction with translation initiation factors, additionally regulating translation. Moreover, 5′-cap has been demonstrated as a recruiter of proteins involved in splicing, polyadenylation and nuclear export. No *cis* element on the gene sequence signals where capping should occur. Instead, capping depends on the activity of the RNA POLII while still in the early stage of elongation. At this transcription stage, POLII has a serine phosphorylation signature in its C-terminal domain (CTD) that elicits activation of the proteins responsible for capping of the nascent pre-mRNA associated with the enzyme. On the other hand, the two other steps of pre-mRNA processing, splicing and polyadenylation, depend on gene *cis* elements to proceed, as well as different POLII CTD serine phosphorylation signatures. In this section, we will review the structure of exons and introns associated with splicing, as well as the *cis* elements guiding polyadenylation and transcription termination.

4.2.4.1 Exons

In 1977, work on the Adenovirus-2 transcription disclosed that its early expression genes shared a common 5′ end in the mRNA and the gene CDS was at discontinuous genomic positions. As adenovirus-2 infects eukaryotic cells, these data were the first evidence that eukaryotic cells should employ splicing to align the CDS in mature mRNA, and awarded the 1993 Nobel Prize in Medicine or Physiology to Phillip A. Sharp and Richard J. Roberts for their discoveries on ‘split genes’. Protein-coding genes from eukaryotic genomes were then immediately demonstrated to have noncoding sequences intervening the CDS on the gene DNA. Although eukaryotic species have a subset of uninterrupted genes, in multicellular species a majority of protein-coding genes is interrupted by <u>int</u>ervening sequences known as <u>int</u>rons, and splicing of the primary transcript must take place to remove introns, consecutively ligate exons and finally display the CDS in a continuous way in the mature mRNA (Fig. 4.2a). In average, humans have eight introns per protein-coding gene. The <u>ex</u>pressed part of the gene, present in the mature mRNA, has been termed <u>ex</u>ons, and may contain untranslated sequences (UTR) besides the CDS. The 5’-UTR extends from the 5′ end to the nucleotide before the translational start codon, whereas the 3’-UTR lies from the nucleotide after the translation termination codon to immediately before the poly-A tail (Box 4.1; Fig. 4.2a and d).

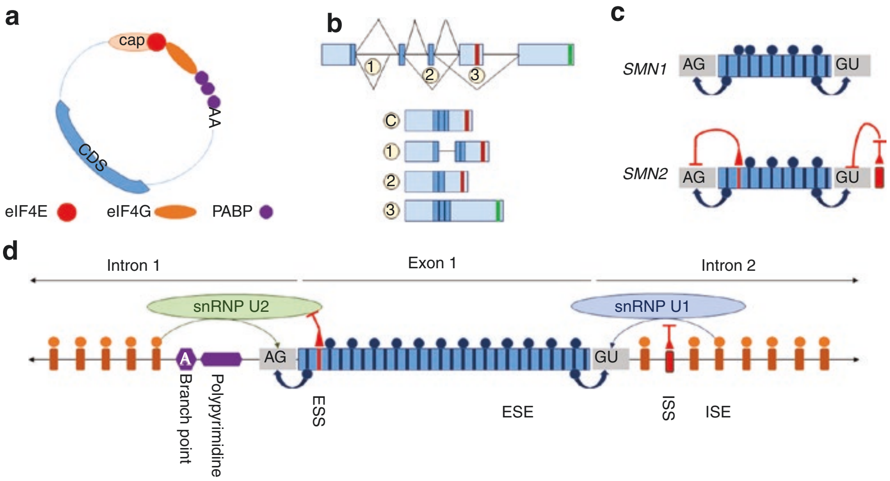


Fig. 4.5 Pre-mRNA processing. (**a**) Eukaryotic mature mRNA adopts a closed configuration through eIF4E binding to cap and eIF4G, which associates with PABP that coats the poly-A tail. (**b**) Alternative splicing may result in intron retention (1); partial intron retention represented); exon skipping (2), choice between mutually exclusive exons (3); in this case with alternative polyadenylation signals—red and green bars). (**c**) Diagram representing *SMN1* and *SMN2* exon 7 with exon definition for *SMN1* and ESS and ISS represented in red repressing *SMN2* exon 7 definition. (**d**) A diagram representing an exon (exon 2) flanked by two introns (showing the 3′ end of intron 1 and 5′ end of intron 2). The intron canonical sites are presented (AG is the acceptor site of the upstream intron and GU is the donor site of downstream intron). The branch point and polyprimidine tract are shown upstream of intron 1 acceptor site. Exonic splicing enhancers (ESE) are indicated as several blue bars in association with SR proteins (small blue circles), and one silencer as a red exonic bar represses splice site choice. Intronic splicing enhancers reinforce exon definition whereas silencers (one indicated) repress exon inclusion. ISE are recognized by hnRNPs (orange circles). In a situation of exon definition, the exon is flanked by both snRNPs U2 upstream and U1 downstream

The 5′- and 3’-UTRs are important regions of the mRNA regulating its stability and translation. These two processes, translation and control of mRNA half-life (ultimately leading to mRNA degradation), are intrinsically related as the eukaryotic translation initiation factor (eIF) 4E (eiF4E) binds to the 5′ cap and interacts with eiF4G, which associates with proteins that coat the poly-A tail (poly-A-binding proteins, PABP; Fig. 4.5a). The closed configuration protects the mRNA ends against degradation by exonucleases and establish the *trans*-acting factors eIF4E, eIF4G and PABP as integrators of translation initiation regulation. This integration is possible owing to the capability of eIF4E, eIF4G or PABP binding to other *trans* factors (RNA-binding proteins, RNA-induced silencing complex RISC, and other indirectly interacting proteins) associated with *cis* elements, mostly in the UTRs, specific for RNA-binding proteins or microRNAs (see Chap. 5).

Analyses of the human genome (Assembly GRCh38.p13 / 109) in the Piovesan et al. [17] report show that although the number of coding exons (151,285) corresponds to 95% of total non-redundant exons (159,652) of protein-coding genes, the total lengths of exon coding (CDS; 25,840,698 bp) and non-coding

(UTRs; 59,281,518 bp) sequences considerably differ. Exon coding sequence length represents 0.8% of the human genome whereas total exon length consists of nearly 2% of the genome sequence. Therefore, as previously discussed, the CDS comprehends a very small fraction of the human genome, and more than half of total human exon sequence is composed of UTRs that may harbor *cis* elements for the regulation of mRNA translation and stability.

In eukaryotes, the translation start codon might be few hundred base pairs downstream of the 5′ cap, making the 43S ribosome translation pre-initiation complex scan the 5’ UTR until the tRNA$^{\mathrm{Met}}$ anticodon base-pairs with the mRNA start codon. This mechanism classifies as cap-dependent because it initiates due to protein-protein interactions among eIFs, notably including eIF4E bound to the 5′ cap. By contrast, an alternative translation initiation mechanism in eukaryotes that is reminiscent of the one functional in bacteria allows the ribosome be assembled directly onto the translational start codon with no need to scan the 5’-UTR. This cap-independent translation initiation mechanism is possible owing to internal ribosome entry sites (IRES), secondary hairpin-like structures adopted by mRNA immediately upstream of the translation start codon, not associated with the 5′-cap complex. IRES-dependent translation operates with higher frequency under cellular stress. Several studies have demonstrated the requirement of cellular RNA-binding proteins, IRES *trans*-acting factors, which might act as RNA chaperones possibly maintaining IRES structure required for the efficient assembly of a pre-initiation complex and recruitment of ribosomes. For instance, IRESs have been reported within the *TP53* full-length mRNA as well as *TP53* variable transcripts for the Δ133p53 and Δ160p53s isoforms. A naturally occurring single nucleotide variant (SNV) in *TP53* 5’ UTR coincident with the full-length mRNA IRES affects IRES binding domain for polypyrimidine-tract binding protein (PTB), thus reducing *TP53* cap-independent translation in steady-state as well as in G2-M checkpoint, and upon DNA-damaging stress and oncogenic insult [15].

In general, the amino acid sequence or quantity of the protein produced by translation may be affected if the CDS results from alternative splicing and/or RNA circularization (see Sect. 4.2.4.2) or if it has alternative translation start codons or alternative ORFs. Alternative translation start codons may result from transcription at alternative promoters, alternative splicing or internal start codon in association with IRES. These processes thus lead to the expression of alternative forms of the protein, known as isoforms (Box 4.2).

Alternative ORFs have been defined as the coding sequence with a translation start codon within any reading frame of either long ncRNAs (see Chap. 5) or known coding mRNAs (either in UTRs or overlapping the CDS). Large-scale ribosome profiling identified widespread translation events outside the annotated CDSs. Part of it was observed upstream or downstream of the annotated CDS, respectively in 5′- or 3’ UTRs, named upstream ORF (uORF) or downstream ORF (dORF), generally with short lengths. DNA variants creating or suppressing a uORF lead to a decrease or increase in the downstream canonical protein expression, respectively, permitting to hypothesize that uORF translation is a regulatory mechanism that may affect translation of the downstream main CDS. In yeast, uORF regulatory roles have been widely demonstrated in genes encoding proteins with survival roles during cell stress. In these cases, the major CDS is the second cistron, and undergoes cap-dependent translation initiation due to an unusual mechanism that keeps the affinity of the 40S ribosome subunit to the mRNA. Additionally, human transcribed polymorphic trinucleotide repeats involved in dynamic mutations in neurologic diseases may be contained within alternative uORFs that are indeed translated upstream the leading CDS (see Chap. 6). However, most alternative ORFs are not yet annotated because of lack of experimental evidence, and their absence in databases precludes their detection by standard proteomic methods [18].

The observation of active minor ORFs in mRNA with a major CDS has brought to light that polycistronic mRNAs is not exclusive of bacteria. Putatively polycistronic mRNAs have been described containing unannotated ORFs in alternative frames in transcripts with one annotated CDS, partially overlapping or completely nested within it. Genome annotation lays the basis for molecular genetic analysis. If in one reading frame a CDS SNV is synonymous, it may indeed be nonsynonymous in a nested frameshifted alternative ORF. Thus functional assessment is necessary to extend the annotation to minor ORFs. Human bicistronic mRNAs have been identified producing two distinct, stable proteins that may physically or functionally interact. As the 3′-most cistron is likely to have translation initiated at IRES, the annotation of nested and overlapping ORFs should thus reflect both features, alternative ORFs and IRES [18].

4.2.4.2 Introns and Splicing

Introns have been classified according to their mechanism of splicing. Nuclear introns employ the spliceosome for splicing whereas introns of types I and II are self-spliced by RNA catalysis. Self-spliced introns are found in some protist, fungi and plant mitochondria DNA, green algae and plant chloroplast DNA and in bacteria. Here we will concentrate in nuclear introns that employ the spliceosome composed of the small nuclear RNAs (snRNAs) U1, U2, U4, U5 and U6, the so-called U2 spliceosome in opposition to the U12 spliceosome that uses snRNA U12 instead of U2 for splicing a small subset of nuclear introns.

The U2 spliceosome depends on U1 and U2 snRNA ribonucleoprotein particles (snRNP) to identify introns in nascent pre-mRNAs by interaction with the conserved canonical splice sites in intron ends: the dinucleotides GU at intron 5′ end (donor splice site) and AG at its 3′ end (acceptor splice site), the branch point adenine and the polypyrimidine tract just upstream of the acceptor site (Fig. 4.5d). This kind of intron (GT-AG) corresponds to nearly 98% of nuclear introns. Basically, upon the interaction of U1 and U2 snRNPs with intron RNA, the U4-U5-U6 snRNPs associate with the former snRNPs, change the configuration of the inactive spliceosome, allowing the exit of U1 and U4 snRNPs and bringing intron ends to proximity in a large complex (the active spliceosome) that will perform splicing catalysis by two transesterification reactions. The expected output for constitutive splicing is the mature mRNA lacking introns and displaying all exons consecutively ligated

(Box 4.2 and Fig. 4.2a). For details of pre-mRNA processing, the reader may refer to review articles [3, 19].

The *cis* splicing (splicing of two RNA segments derived from the same primary transcript) mechanism summarized above has evolved allowing efficient removal by the spliceosome catalysis of even very long introns, as found in mammals (Fig. 4.1b). Consequently, introns may evolutionarily 'gain' sequences without affecting the splicing efficiency unless if the splicing sites or regulatory elements are lost (see below). Therefore, introns commonly harbor repeat elements of various classes (see Chaps. 6, 8 and 9), as exemplified in the *TP53* gene introns (Fig. 4.4, bottom part), in which lies the majority of sequence repeats of this gene, in addition to the 3' UTR.

Optimal exon length close to an average length (150 nucleotides in human), transcription elongation rate, and chromatin configuration all contribute to the combinatorial effect for constitutive splicing. The limited length and degenerate nature of *cis* elements that allow U1 and U2 snRNPs recognize the intron RNA would render mRNA susceptible to a larger number of errors if *cis*-regulatory elements were not in place. Exon definition in nascent pre-mRNA is an essential step before splicing takes place. U2 snRNP assembly in the 3′ end of the upstream intron and U1 snRNP association with the 5′ splice site in the downstream intron make a configuration known as cross-exon, which is modulated by several *cis*-acting elements, besides the 5′ and 3′ splice sites, the branch point and the polypyrimidine tract (Fig. 4.5d). Serine-arginine (SR)-rich proteins are *trans* factors recognized as splicing activators when bound to exon splicing enhancers (ESE) though they can act as repressor *trans* factors if bound to intron splicing silencers (ISS). Likewise, intron splicing enhancers (ISE) add to exon definition by recruiting *trans* factors belonging to the heterogeneous nuclear ribonucleoprotein (hnRNP) family. HnRNP in association with exon splicing silencers (ESS) tend to repress exon definition (Fig. 4.5d) [3].

The suboptimal representation of splicing enhancer elements (ESE and ISE) and/or the occurrence of silencer elements (ESS and ISS) may decrease the strength of canonical splice sites and modify the constitutive splicing output, producing alternative splicing. Alternative splicing outcome has been classified as partial or full exon skipping or intron retention, and mutually exclusive exons (see Fig. 4.5b and Sect. 4.2.4.4). In addition, partial exon skipping or intron retention may be the result of a cryptic splice site selection instead of the canonical 5′ donor or 3′ acceptor sites within the respective exon or intron. If alternative splicing causes in-frame alterations of the CDS, the consequent change in protein sequence may increase protein isoforms and proteome diversity. As more alternative transcripts affecting protein sequence and domains are annotated for specific genes, the larger is the effect on proteome diversity (Box 4.2 and Fig. 4.1c) correlating to increasing phenotypic complexity along the phylogenetic scale. By contrast, alternative splicing affecting the CDS may shift the reading frame, creating premature translational termination codons, which can elicit the nonsense-mediated decay (NMD) of the mRNA. Some naturally occurring, regulated alternative exons have been termed poison exons because their inclusion in mRNA introduces a premature translation termination codon that can be involved in auto-regulation of the gene's expression, as they may elicit NMD [3].

The phylogenetic increase observed for the genome's number of introns (intronization) may have been due to species-associated DNA variants creating splice canonical sites or modifying *cis*-regulatory enhancers (see Sect. 4.2.4.3) and silencers, and by insertion of transposable elements or sequence duplications. Under an evolutionary perspective, transposable elements inserted within a gene sequence can affect the interaction of the pre-mRNA with RNA-binding proteins in the spliceosome, leading to intronization or, conversely, the acquisition of exon features (exonization) as reported for Alu repeats (See Chap. 8).

Recent high-throughput sequencing of cDNA from non-polyadenylated RNA transcripts has identified circular RNAs (circRNAs) in various eukaryotic species. A large subset of circRNAs has been demonstrated to result from back-splicing of pre-mRNA, in which the 5′ splice site (donor site) is downstream of the 3′ splice site (acceptor site). Functional assessment of circRNAs show that some of them can be translated into functional peptides, others constitute a hybridization platform for microRNA consequently reducing their availability (sponge effect) and, still, a few that retain an intron between exons in the circular molecule can enhance the transcription of their parental genes. In summary, back-splicing, a non-canonical form of splicing, may produce circular molecules of RNA that are not polyadenylated, but stable, and may act as regulatory scaffolds for microRNA, gene transcription or translation [19].

Nearly 20% of pathogenic DNA variants causing human genetic diseases are believed to be splicing variants, at least 70% of them directly affecting canonical splice sites. In rare occasions, synonymous variants or variants annotated as missense or nonsense may affect ESEs or create canonical splice sites on exon borders. However, their effect must be assessed on mRNA or by a minigene approach, in a research setting. Approximately 3 to 5% of all disease-causing variants should lie deeply in introns, having been hardly identified either because intron internal sequences are normally not included in NGS DNA libraries in a genetic testing setting (see Chaps. 1 and 3) or the effects of rare deep intronic variants on splicing should not have been demonstrated. Deep intronic DNA variants may activate cryptic splice sites and include pseudoexons in the mature mRNA, changing the translational reading frame or not. Many mRNA products modified by deep intronic splicing variants likely undergo NMD based on production of premature translation stop codons, as seen for part of alternative splicing products [3].

Quantitative trait locus (QTL) is a locus that correlates to the quantitative variation of a specific trait in a population. The analyses of transcriptomes have expanded the term QTL to expression QTL (eQTL) meaning the amount of mRNA related to a gene locus and its DNA variants, *e.g.* SNVs on gene promoter. More recent analyses of transcriptome (RNA-Seq) data sets adapted the molecular eQTL approach to assess transcript variants (splicing QTL or sQTL) due to alternative splicing or SNVs altering *cis*-acting elements regulating pre-mRNA splicing. In sQTL, the trait can be, for instance, the estimation of transcriptome deep sequencing reads by exon or reads spanning exon junctions, when assessing exon exclusion [3].

Rare SNVs are abundant in human genomes and include splicing SNVs of uncertain clinical significance. The Genotype-Tissue Expression (GTEx, The BROAD Institute of MIT and Harvard University, Cambridge, MA) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. The combined analyses of high throughput genomic and transcriptomic data from GTEX allow to assess how rare genetic variants contribute to extreme patterns in alternative splicing by disrupting or creating a splicing consensus sequence. SNV-based creation of splice site may include cryptic exons in mature mRNA, disrupting or maintaining the CDS reading frame (https://www.gtexportal.org).

4.2.4.3 The *SMN1* and *SMN2* Genes: A Genetic Case Study on Splicing

Spinal muscular atrophy (SMA) is a rare autosomal recessive, neuromuscular disorder characterized by degeneration of the anterior horn cells in the spinal cord and the brainstem nuclei, resulting in progressive muscle weakness and atrophy. The weakness is symmetric, more proximal than distal, and progressive; and the onset varies from prenatal stage to adulthood. It classifies in a five-degree spectrum of neuromuscular phenotypes at presentation related to predicted possible outcomes. Supportive measures, adequate nutrition, respiratory assistance, and preventing complications of muscle weakness have composed the clinical management until the recent association of targeted treatments.

Homozygous loss of function of the *SMN1* gene (Survival of motor neuron 1; OMIM 600354; 5q13.2) causes SMA. Individuals with SMA are either homozygous for a deletion encompassing at least *SMN1* exon 7 or are compound heterozygotes for such deletion and an inactivating pathogenic DNA variant in *SMN1*.

SMN2 (OMIM 601627; 5q13.2) is a gene paralogous to *SMN1*, showing a high degree of nucleotide identity (99%). Both genes´ CDSs encode 294-amino acid RNA-binding proteins, survival motor neuron 1 and 2 (SMN1 and SMN2), required for efficient assembly of specific ribonucleoprotein complexes. Whereas *SMN1* produces a full-length SMN1 protein necessary for lower motor neuron function, *SMN2* expresses SMN2 lacking the segment encoded by exon 7 rendering it a less stable truncated protein. Therefore, upon biallelic inactivation of *SMN1*, *SMN2* cannot fully compensate for loss of the SMN1 protein in motor neurons.

The lack of inclusion of exon 7 in mature *SMN2* mRNA is in part due to a substitution in the 5′ end of exon 7 that abolishes an ESE creating an ESS that impairs snRNP U2 efficient exon 7 definition (Fig. 4.5c). It had been reported that SNVs in exon 7 of *SMN2* may act as SMA modifier resulting in a milder clinical phenotype. Therefore, therapeutic strategies that enhance the expression of either full-length protein, SMN1 or SMN2, have been developed. In the last three years, three drugs have been approved for SMA patients. Nusinersen (Spinraza™, Biogen, Switzerland) is a pharmaceutical drug consisting of an antisense oligonucleotide targeting the *SMN2* pre-mRNA silencer in intron 7, promoting *SMN2* exon-7 inclusion, and thus increasing the expression of the full-length SMN2 protein. Evrysdi (Risdiplam™, Roche, Switzerland) is a small molecule that promotes *SMN2* exon 7 inclusion in mRNA with high specificity. Although its mechanism is not fully understood, it is believed to bind to exon 7 *cis* elements displacing *trans*-acting factors that repress exon inclusion. Zolgensma (Onasemnogene abeparvovec-xioi, Novartis, Switzerland) is an adeno-associated virus 9 (AAV9)-based gene delivery for reposition of the SMN1 protein [20].

4.2.4.4 Polyadenylation Signal: A *cis* Element on Transcript's Last Exon

Eukaryotic mRNA polyadenylation is the third and final step in processing of the primary transcript. It depends on the recognition of the polyadenylation signal on the last exon of the transcript (Fig. 4.2a), which has a consensus 5' AAUAAA 3′ on nascent pre-mRNA. Before polymerization of the poly-adenine (poly-A) tail the mRNA must be cleaved nearly 20 ribonucleotides downstream of the polyadenylation signal (Box 4.2). This cleavage displaces the already processed mRNA off POLII that remains associated with an uncapped transcript, reducing its polymerization processivity, an accepted model to terminate transcription. On the other hand, on the 3′ end of the cleaved pre-mRNA, a polyadenylate polymerase actively elongates it through addition of a series of nearly 200 adenines. The growing poly-A tail becomes coated by PABP, which is pivotal in allowing the closed configuration of the mature 5′-capped, 3′-polyadenylated mRNA by association of their ends with translation initiation factors (Fig. 4.5a). Therefore, as the 5′ cap, the poly-A tail is important for stabilization of the mRNA and its translation regulation. PABP protects the mRNA against the action of 3′-to-5′ exonucleases, avoiding deadenylation of the poly-A tail.

The length of the poly-A tail is directly related to the mRNA half-life. The UTRs, mostly the 3'-UTR of the mRNA, also play major roles in regulating mRNA stability as they may harbor *cis* elements with regulatory activities over enzymes that control the stability of 5′ cap and 3′ poly-A. Activated decapping or deadenylation leads to rapid degradation of the mRNA due to exposure of its 5′ or 3′ ends, respectively. Another way to differentially control the mRNA stability or translation is by modifying the length and composition of its 3' UTR. If the UTR is composed by more than one exon (meaning that the translation stop codon is not on the last exon) or the gene's last intron has alternative 3′ splice sites, alternative splicing may affect the 3'-UTR. When two polyadenylation sites lie in mutually exclusive exons splicing should lead to conditional inclusion of one of them in mature mRNA. In Fig. 4.4, *TP53* transcripts initiated at intron 1 alternative promoters underwent alternative splicing employing one of two mutually exclusive exons downstream of the constitutive 3′ end-most exon employed by full-length transcripts (Fig. 4.4, red arrowheads on *TP53* transcripts). Moreover, the gene may have alternative poly-A sites or additional mildly degenerate poly-A sites that can be recognized under certain conditions when *trans*-acting factors hinder the constitutive site in the nascent pre-mRNA. Alternative polyadenylation altering the length of the 3'-UTR of cancer-associated genes may modulate their expression by affecting *cis* elements for miRNA and RNA-binding proteins. Finally, all *cis* elements mentioned in this section are part of the gene sequence, as they are transcribed and recognized on the pre-mRNA or the mature mRNA 3' UTR, regulating, respectively, polyadenylation or mRNA stability and translation.

4.2.5 The ENCODE Project

The ENCyclopedia Of DNA Elements (ENCODE) project was launched by the ENCODE project consortium in 2003 upon the near completion of the human genome sequencing. At long term, it aims at developing a comprehensive map of functional elements in the human genome. This map will include but will not be limited to the full annotation of protein-coding and non-coding genes, variable transcripts, epigenetic marks (histone modifications and DNA methylation) as well as sites of open chromatin accessible to DNAse I digestion (Fig. 4.3a) or to specific transcription factors, inter- and intra-chromosomal interactions, RNA-protein binding sites, and comparative genomics. For the latter, the ENCODE project also assessed genomic data of three other model organisms (the free-living soil worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and the house mouse *Mus musculus*). The ENCODE consortium comprehends laboratory and computational scientists that follow standardized protocols to contribute with datasets that become available at the project portal (https://encodeproject.org). The site https://encodeproject.org/about/data-acess lists alternative providers of genome browsers for analysis of limited genomic regions from the ENCODE data.

In the ENCODE pilot phase 1 (2003–2007), nearly 1% of the human genome (30 Mb) was selected for initial analyses and testing of technological strategies. Phase-2 (2008–2012) interrogated the whole genome and transcriptome by implementing high-throughput sequencing-based technologies (see Chaps. 3 and 5). The ENCODE project phase 3 (2013–2016) added RNA binding experimental approaches and the three-dimensional (3D) organization of chromatin assessed by chromatin interaction analysis. In particular, phase 3 allowed for mapping of transcribed regions and transcript isoforms, regions of transcripts recognized by RNA-binding proteins, transcription factor binding sites, and regions that harbor specific histone modifications, open chromatin, and 3D chromatin interactions. The ENCODE project will conclude its fourth phase in 2021, and its goal is to expand the phase-3 analyses into more cell types and tissues to attempt to gain a more integrated view of mapped open chromatin regions and the transcription factors that bind to these sequences [21]. Finally, so far along the four phases of the ENCODE project, the expression of non-coding RNAs (ncRNA) has been disclosed with high occurrence in the human genome, as exemplified in Section 4.2 and in Chap. 5 of this book.

4.3 Protein-Coding Gene Families

The Reference Sequence (RefSeq) database at the National Center for Biotechnology Information (Bethesda, MD, USA; https://www.ncbi.nlm.nih.gov/refseq) is a non-redundant collection of richly annotated DNA, RNA, and protein sequences from taxonomically diverse organisms. Each RefSeq accession number represents a single, unique, full sequence of a gene or a naturally occurring mRNA or protein molecule from one organism. The RefSeq database has been developed and curated by NCBI staff and collaborators after the sequencing of many model organism genomes. It has the goal to provide a comprehensive, standard dataset that represents sequence information for a species. It has been built using data from public archival sequence databases (eg. Genbank) that could present replicate entries and sequences that were incomplete, redundant or with ambiguously sequenced bases. The rapid growth of the RefSeq database reflects the inclusion of non-model organisms for the last two decades. For instance, on June 30, 2003, 39 vertebrate mammalian genomes had been submitted to curation and annotation by the RefSeq approach, whereas on November 2, 2020, that number had increased to 573. The increasing annotation of protein-coding genes in the human genome by RefSeq and initiatives of the ENCODE project approaches a more reliable estimate of the number of those genes, discarding duplicate records and sorting out pseudogenes. Therefore, while sequence curation will be still reviewed and necessary, that number will be continuously updated in the most recent assembly of the respective genome sequence. One should take into consideration that the annotation systems may differ among distinct initiatives leading to slight differences in total numbers and classifications, as seen for the annotated human protein-coding genes that sum up 19,405 and 19,945, respectively for NCBI human genome assembly GRCh38.p13 (annotation release 109; Fig. 4.1a) and GENCODE (V35; see below and Chap. 1).

The GENCODE project (https://www.gencodegenes.org), a branch of the ENCODE project funded by the National Human Genome Research Institute (NHGRI, Bethesda, MD) and the European Molecular Biology Laboratory (https://www.embl.org), aims to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, by manual curation, computational analysis and targeted experimental approaches. In this chapter we have examples retrieved from both initiatives, the NCBI Genome (Fig. 4.1) and RefSeq (Fig. 4.4 and Sect. 4.3.3) databases and the ENCODE and GENCODE projects (Figs. 4.4 and 4.6). Apart from the 19,945 annotated protein-coding genes of the human genome by the GENCODE initiative (GENCODE V35, accessed in November, 2020), 14,767 pseudogenes have their sequences annotated and roughly classify into processed (nearly 72%), unprocessed (~24%), unitary (~2%), polymorphic (less than 0.5%) pseudogenes, and miscellaneous or other (~1.5%).

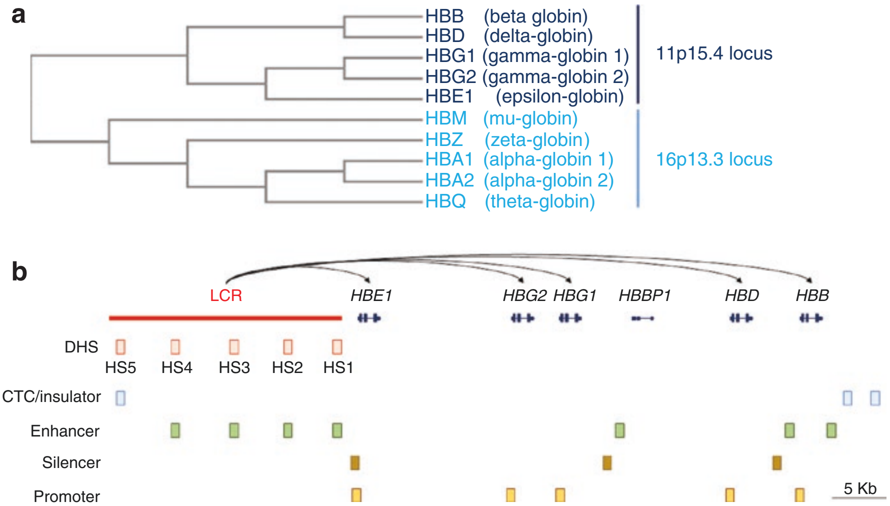


Fig. 4.6 (**a**) A phylogenetic tree (cladogram type) obtained upon multiple alignment of protein sequences of human globin paralogs encoded by the gene clusters in chromosomes 11p15.4 (beta-globin cluster; dark blue) or 16p13.3 (alpha-globin cluster, light blue). Alignment and cladogram were obtained using the CLUSTAL Omega program (https://www.ebi.ac.uk/Tools/msa/clustalo/) and the RefSeq sequences retrieved at NCBI Entrez (https://www.ncbi.nlm.nih.gov/protein/) for HBB (NP_000509.1), HBD (NP_000510.1), HBG1 (NP_000550.1), HBG2 (NP_000175.1), HBE (NP_005321.1), HBM (NP_001003938.1), HBZ (NP_005323), HBA1 (NP_000549.1), HBA2 (NP_000508.1) and HBQ (NP_005322.1). (**b**) A diagram view of a small part of human chromosome 11 band p15.4 (11p15.4), spanning 75 kb (5,225,001 to 5,300,000), displaying the cluster of beta-globin genes and its locus control region (LCR). The coding genes are indicated by acronyms according to the proteins they encode seen on upper panel of (**a**), and *HBBP1* stands for *HBB* pseudogene 1. Arrows indicate the direction of regulatory control based on LCR towards each individual gene. The position of some DNAse hypersensitive sites (HS), CCTF binding sites characterizing insulators, as well as enhancer, silencer, and promoter elements are indicated. Drawing was made to scale according to genomic data retrieved at the University of California in Santa Cruz (UCSC) genome browser accessed in November, 2020. Scale bar: 5 kb

4.3.1 Pseudogenes

A pseudogene is any genome sequence identified by similarity to another gene, but defective due to evolutionarily accumulated inactivating mutations. Pseudogenes belong to both protein-coding and non-coding gene families and their sequences intersperse in the genome. Pseudogene classification into processed and unprocessed is based on their mechanisms of origin. Processed pseudogenes derive from retrotransposition of processed mRNAs, whereas unprocessed pseudogenes arise from segmental duplication of genomic DNA. Retrotransposition implies reverse transcription of the mRNA of a functional gene and the integration of this complementary DNA (cDNA) copy into a site of the genome that can be far from the locus of the functional gene. Consequently, if the original functional gene contained introns, its processed pseudogene will likely not. Additionally, processed pseudogenes may retain a replica of the poly-A tail from the mRNA that served as template for its cDNA copy maintaining poly-adenine features in its 3′ end. Of note, direct repeats may be reported in the site of genome integration of the cDNA. Diverse mechanisms may generate segmental genomic duplications (see Chaps. 8 and 9) that originate unprocessed pseudogenes: apparently most frequently tandem duplications by unequal crossover, as well as partial genome duplication through nondisjunction, transpositions involving transposable elements, and duplications occurring after rearrangements and subsequent repair of staggered breaks. Sequence alterations can inactivate the gene promoter motifs and *cis*-regulatory elements of unprocessed pseudogenes abolishing its transcription. Since pseudogenes are generally not submitted to positive selection, the CDS derived from the original gene tends to accumulate mutations frequently creating premature translation termination codons that extinguish or truncate the protein product [22].

Unitary pseudogenes are a minor class of pseudogenes that are not processed, formed without duplication by inactivation of a single original gene through DNA variant occurrence that eliminates the functional copy of the gene. The rare polymorphic pseudogenes have mutations in the reference genome but are intact in some individuals [22].

Protein-coding gene CDS and regulatory sequences undergo strong selective pressure that is essential to the development and function of the organism and species survival. By contrast, most pseudogenes evolve neutrally, making them an ideal alternative for the study of genome evolution. Since nearly three quarters of human pseudogenes are of the processed type, their contribution to the human genome size is limited as they lack introns. Multiple lineage sequence analyses and observations that processed pseudogenes prevail largely across the human genome in areas of low recombination rates favor the hypothesis that this class of pseudogenes are lineage-specific and in humans probably resulted from a retrotranspositional burst at the dawn of the primate lineage that occurred ~40 millions of years ago. The predominance of human processed pseudogenes belonging to large families of protein-coding genes with high expression levels also corroborates that hypothesis [22].

Since a pseudogene is considered a defective copy of a gene, database annotated pseudogenes are not taken into account for the biological and functional impact of the genome. However, the duality between functional and non-functional aspects of a gene based only on the neutral effect of the pseudogene CDS has been questioned in the past decade, specifically for purposes of database annotation of transcribed pseudogenes. Transcriptome analyses have disclosed that transcription is pervasive in the human genome involving diverse classes of DNA elements discussed in this book (see Chaps. 5 to 9), including at least 15% of pseudogenes. Regulatory elements may have become evolutionarily associated with processed pseudogenes or reside within their sites of genome integration, activating their transcription. Pseudogenes frequently exhibit high sequence similarity with the ancestral (commonly referred to as parental) gene. It has thus been hypothesized and demonstrated in a few cases that antisense pseudogene transcripts could regulate levels of expression of sense parental gene by a direct interference by duplex formation between antisense pseudogene and parental sense gene mRNAs or generation of short interfering RNAs (siRNA). In addition, parental gene expression could be modulated by sense pseudogene mRNA through endogenous competition for positive or negative factors that control stabilization of both pseudogene and parental gene mRNAs or directly interfering as long antisense RNA on the parental gene chromatin regulating its transcription [23].

Some genes with characteristics of origin by retrotransposition ('retrogenes') have been recognized as functional protein-coding genes rather than pseudogenes across species. An additional group of genes, comprehending nearly 60 members, show evidence to have been formed by retrotransposition or gene duplication since hominid divergence from other primates, constituting new human-specific or hominid-specific protein-coding genes. This is of great interest to understand cerebral cortex increase in size and complexity related to significant acquisition of cognitive functions in the human species. On the other hand, genes may be classified and annotated as pseudogenes but later demonstrated to express functional truncated peptides, being in fact a functional protein-coding gene [23].

The human-specific *NOTCH2NLA*, *NOTCH2NLB* and *NOTCH2NLC* (Notch 2 receptor N-terminal like A, B and C) genes in chromosome 1 appear to be the result of segmental duplications of the ancestral *NOTCH2* gene. Their genomic organization consists of the first four exons and introns conserved in *NOTCH2* and a fifth exon apparently derived by exonization of an intronic segment of *NOTCH2*. All NOTCH2NL paralogs are predicted to contain an ORF encoding a protein homologous to NOTCH2 but truncated as they include only the N-terminal region of NOTCH2 extracellular domain. When compared to *NOTCH2* gene structure of 34 exons and 33 introns, NOTCH2NL paralogs would have gene annotation trends for their classification as unprocessed pseudogenes. However, recent studies have disclosed that the three NOTCH2NL paralogs express peptides in embryonic telencephalic vesicles that are secreted and enhance NOTCH signaling between the seventh and ninth human gestational weeks during early organogenesis of the cerebral cortex. As their overexpression delays cerebro-cortical neurogenesis, *NOTCH2NLB* was demonstrated to have effects on clonal expansion of human cortical progenitors by directly affecting the self-renewal of the progenitor cell pool before neuronal differentiation. Additionally, *NOTCH2NLA* and *NOTCH2NLB* serve as breakpoints in the 1q21.1 copy number variation (deletion/duplication; see Chap. 9) associated with clinical phenotypes of microcephaly/macrocephaly, intellectual disability and behavioral deficits including autism spectrum features [23].

4.3.2 Protein Domains as Evolutionary Modules

Protein domains are structural, functional and evolutionary units, consisting of amino acid sequences of limited length (50–250 amino acids) that generally fold independently producing a specific protein tertiary structure stable by itself. Protein domains represent a module able to function alone or together with partner domains on the same protein chain contributing to its specific functions [24].

The combination of different domains in a single chain gives rise to a large variety of proteins and generates functional diversity. Domain architecture or domain organization refers to the order of all domains in a protein chain from N- to C-terminus, directly related to the translation of the CDS of the encoding gene from 5′ to 3′ end. Thus, domain architecture refers to multi-domain proteins with complex functions.

The central axis of comparative genomics is sequence comparison by multiple alignment generating phylogenetic trees. For domain sequence comparison the similarity is higher at the protein level than the nucleotide (CDS) level as the genetic code is degenerate mostly due to variation in codon third base. Nevertheless, for a few protein domain comparisons the homology will only become evident if the tertiary structure of the protein domain is experimentally obtained. Homology refers to structures or sequences that evolved from a common ancestral structure or sequence; herein only referred to sequences. Sequence homology can classify as orthologous or paralogous sequences. Orthologous sequences in two organisms are homologs that evolved from the same ancestral sequence by modification, reflecting the organism evolution and speciation. Conversely, paralogous genes are homologs in the same organism generated by duplication of an ancestral gene. Consequently, genes related by ancestry in the same species constitute a gene family, also known as protein family for protein-coding genes.

Domain architecture of multi-domain proteins exhibits a strong correlation with the exon-intron genomic structure of the encoding gene. Each individual domain tends to be encoded by one exon or a combination of exons, and may be considered the translational product of a single evolutionarily mobile CDS module that can undergo position change in the genome, duplication or deletion by mechanisms mostly dependent on genomic recombination. One mechanism, exon shuffling, refers to copy of an exon, or a genomic segment containing few exons, from one gene and transfer into another gene by intron recombination. Exon shuffling has been proposed as an important genetic mechanism sharing exons between non-homologous genes, and driving protein evolution through novel domain acquisition and likely gain of function. On the other hand, domain acquisition may also occur by other forms of non-homologous recombination, retrotransposition mediated by intronic retrovirus repeat units, or gene fusion due to loss of transcriptional signals between adjacent genes. Likewise, protein domain can be lost due to DNA variants modifying the CDS, splice sites, transcriptional and or translational elements [24].

4.3.3 Developmental Expression of the Human Beta-Globin Locus Genes

Human hemoglobin consists of a quaternary protein structure of two alpha- and two beta-globin subunits, known as α2β2 tetramer, having each polypeptide chain a heme unit in association. We will employ the human hemoglobin as an example to discuss in this section the locus of beta-globin genes in chromosome 11 as a model for gene expression control related to human diseases and, in Section 4.4, we will illustrate Mendelian inheritance patterns with examples of genetic heme biosynthesis deficiencies.

The globin gene family has members distributed in clusters in two human chromosomes, the alpha-globin locus at 16p13.3 and the human beta-globin locus at 11p15.4. Human alpha- and beta-globin multigene cluster loci contain three and five functional paralogous genes, respectively, with similar orientation, each originated by duplication of a common ancestor, encoding proteins with related tertiary structures. Amino acid similarities are higher within protein isoforms encoded by genes from the same cluster (Fig. 4.6a).

The five functional beta-globin genes and three alpha-globin genes are arranged in the order of their developmental expression, which is individually controlled by regulatory elements located upstream of each globin gene cluster in chromosomes 11 or 16, the multispecies conserved regions in the alpha-globin locus and the beta-globin LCR. When each one contacts promoters, globin gene transcription may be activated in erythroblast cells (nucleated red blood cells) in a way that along the development each gene becomes uniquely and coordinately activated (Fig. 4.6b; arrows), a process known as hemoglobin switching that meets the changing oxygen demands of the growing embryo or fetus. The earliest embryonic human hemoglobin tetramer (ε2ζ2) is expressed in the yolk sac. At approximately eight weeks of gestation the ε2ζ2 tetramer is gradually replaced by the adult alpha-globin chain and two different fetal beta-like chains, gamma 1 or gamma 2, expressed in fetal liver, spleen, and bone marrow. The α2γ2 tetramer becomes the predominant hemoglobin throughout the remainder of fetal life and, just before birth, the gamma-globin chains are gradually replaced by the adult beta-globin and delta-globin, producing the α2β2 and α2γ2 tetramers in the bone marrow. Six months after birth, nearly 98% of hemoglobin is α2β2, while the α2γ2 tetramer accounts for approximately 2%. As seen, the alpha-like and beta-like globin gene clusters have coordinated programs for differential gene expression. While gene selection switches twice (embryonic to fetal to adult globins) for the beta-like genes, a single switch in the alpha-globin locus shuts down production of zeta-globin early in fetal life [13].

The beta-globin LCR forms physical contacts with active genes in this locus via chromatin looping and contains five DHS (HS1, HS2, HS3, HS4, and HS5; Fig. 4.6b). Four of them (HS1 through HS4) harbor erythroid-specific DNA enhancer motifs recognized by different transcription factors (Fig. 4.6b). The fifth DHS (HS5) contains binding sites for CTCF that function as an insulator having enhancer-blocking activity (Fig. 4.6b). Each globin gene in the beta cluster presents one or two CCAAT boxes, GATA sites for the GATA-binding protein GATA-1 and CACCC-box as promoter-proximal elements, besides a TATA-box in the core promoter. Erythroid-specific long-range interactions have been observed *in vivo* between the active murine beta-globin gene and the LCR, and it appears that GATA-1 activator mediates promoter loops. These long-range interactions of the beta-globin gene were not observed in cells that do not express globin. Conditional knockout of the *Ctcf* gene limited chromatin looping in the mouse beta-globin locus, while ectopic addition of a single CTCF-binding site insulator in the human locus induced the formation of alternate loops that disrupted communication with beta-globin gene promoters [13].

The majority of genetic variants causing hemoglobinopathies are short-range DNA variants within the globin gene CDS, canonical splice sites or the promoter. Nonsynonymous substitutions create hemoglobin variants, such as the sickle variant HbS (a pathogenic DNA variant in the *HBB* gene resulting in the amino acid substitution p.Glu6Val), while DNA alterations causing a quantitative loss of mRNA may give rise to reduced globin chain synthesis and a thalassemia phenotype of microcytic hypochromic anemia. The absence of beta-globin production causes beta-zero-thalassemia, while reduced amounts of detectable beta-globin protein causes beta-plus-thalassemia. Thalassemia clinically classifies in (transfusion-dependent) thalassemia major and thalassemia intermedia (intermediate severity) in homozygotes, and thalassemia minor (asymptomatic) in heterozygotes (carrier state). Infants with thalassemia major may present life-threatening anemia, failure to thrive, and jaundice, and may develop skeletal changes, hepatosplenomegaly and dilated cardiomyopathy that can be in part avoided if a regular blood transfusion program is maintained. If iron chelation therapy is not associated, thalassemia major patients may develop severe complications of iron overload, which in adolescence may include retardation of growth and sexual maturation, besides those related to hemochromatosis (cardiomyopathy and pericarditis, hepatitis and liver cirrhosis, as well as specific gland insufficiencies) [13].

The demonstration that deletions encompassing the β-globin LCR DHS inactivated the full set of globin genes resulting in thalassemia was among the first examples of altered gene regulation as a mechanism of human genetic diseases. In the absence of short-range pathogenic DNA variants in the globin genes, disruption in the linear relationship between the clustered genes and their distant *cis*-regulatory elements, although rare, constitutes a molecular pathophysiology mechanism for thalassemia. Most commonly, loss-of-function pathogenic DNA variants associate with beta-zero-thalassemia, while naturally occurring pathogenic DNA variants in the *HBB* promoter (CACCC and CCAAT boxes and TATA box), UTRs, splice sites off the canonical sites associate with beta-plus-thalassemia due to reduced expression of the gene. By contrast, DNA variants in the CCAAT box associated with the *HBG1* or *HBG2* genes can activate the expression of the silenced gamma-globin genes in the adult causing an unusual benign condition termed hereditary persistence of fetal hemoglobin. Understanding this mechanism has been reasoned as a potential therapeutic intervention for beta-thalassemia patients, as some patients with a mild phenotype of beta-zero thalassemia have at least one of their pathogenic alleles linked to an activating variant in the *HBG2* gene. This has been explained by overlapping sites for the transcriptional activator NF-Y and a repressor of gamma-globin genes. Curiously, under the same reasoning, *HBBP1*, a pseudogene right downstream of the two gamma-globin genes (Fig. 4.6b) is known to enable the dynamic chromatin changes that regulate expression of fetal and adult globin genes during development. Notably, although inhibiting *HBBP1* transcription has no regulatory effect, deletion of this pseudogene reactivates fetal globin expression. *HBBP1* DNA contacts, but not transcription, are required for suppressing the expression of fetal globin genes in adult erythroid cells [13].

4.4 DNA Variants May Occur *de novo* or Be Inherited

DNA alterations may arise as result of DNA polymerase errors during DNA replication in the S-phase of the cell cycle (see Chap. 2) or due to genotoxic effects of physical (e.g., ionizing or ultraviolet radiation), chemical (e.g. oxidative reactive species, aflatoxins) or biological (eg., viruses) agents. A novel germline DNA modification that escapes the action of the DNA repair machinery and segregates to future generations is fixed. DNA variants that occur in germ cells or their progenitors ('gonia' stage) are in the germline and may be transmitted to a whole individual upon fertilization. Similarly, if the variation arises soon after fertilization (post-zygotic stage), most cells of the individual will harbor the DNA alteration. Conversely, DNA alterations that originate in somatic cells will be confined to that individual's particular tissue and, thus, not inherited.

The pipeline for pathogenicity definition of a human DNA variant has been presented in Chap. 3. As proteins are the final product of the expression of protein-coding genes and mRNA the intermediate product, their isolation from tissues and conduction of cell assays pursuing to assess *in vitro* the effects of potential loss-of-function DNA variants make it more amenable to functionally analyze CDS DNA variants in a research setting. However, reporter genes expressed by prokaryotic or eukaryotic vectors have been largely employed to evaluate the biological role of noncoding DNA *cis* elements in regulating transcription, splicing, translation and RNA stability, thus contributing to define the pathogenicity of rare noncoding DNA variants in research.

Porphyrias are disorders that can be genetic or acquired. Genetic porphyrias are caused by pathogenic DNA variants in genes coding for enzymes of the heme (iron protoporphyrin) biosynthetic pathway. In 50% of the cases, the DNA variant is inherited as an autosomal dominant or recessive, or X-linked fashion. In the remainder of cases, the variant is novel in the family having occurred *de novo*. Porphyrin is synthesized in erythroblasts and liver starting in mitochondria by the condensation of glycine and succinyl CoA forming δ–aminolevulinate, which is transferred to the cytosol, where most reactions will take place until protoporphyrinogen is up-taken by mitochondria and heme synthesis finalized (Fig. 4.7). Successive condensation reactions lead to the formation of a tetrapyrrole that then cyclizes forming the porphyrin skeleton. Its side chains undergo chemical modifications to finally react with iron to produce heme [25].

Porphyrias may classify clinically as cutaneous or acute porphyrias. Cutaneous porphyrias are characterized by cutaneous photosensitivity resulting in tingling, burning, pain, and itching, accompanied by swelling and redness upon exposure to sunlight/light. Cutaneous porphyrias include erythropoietic porphyria,

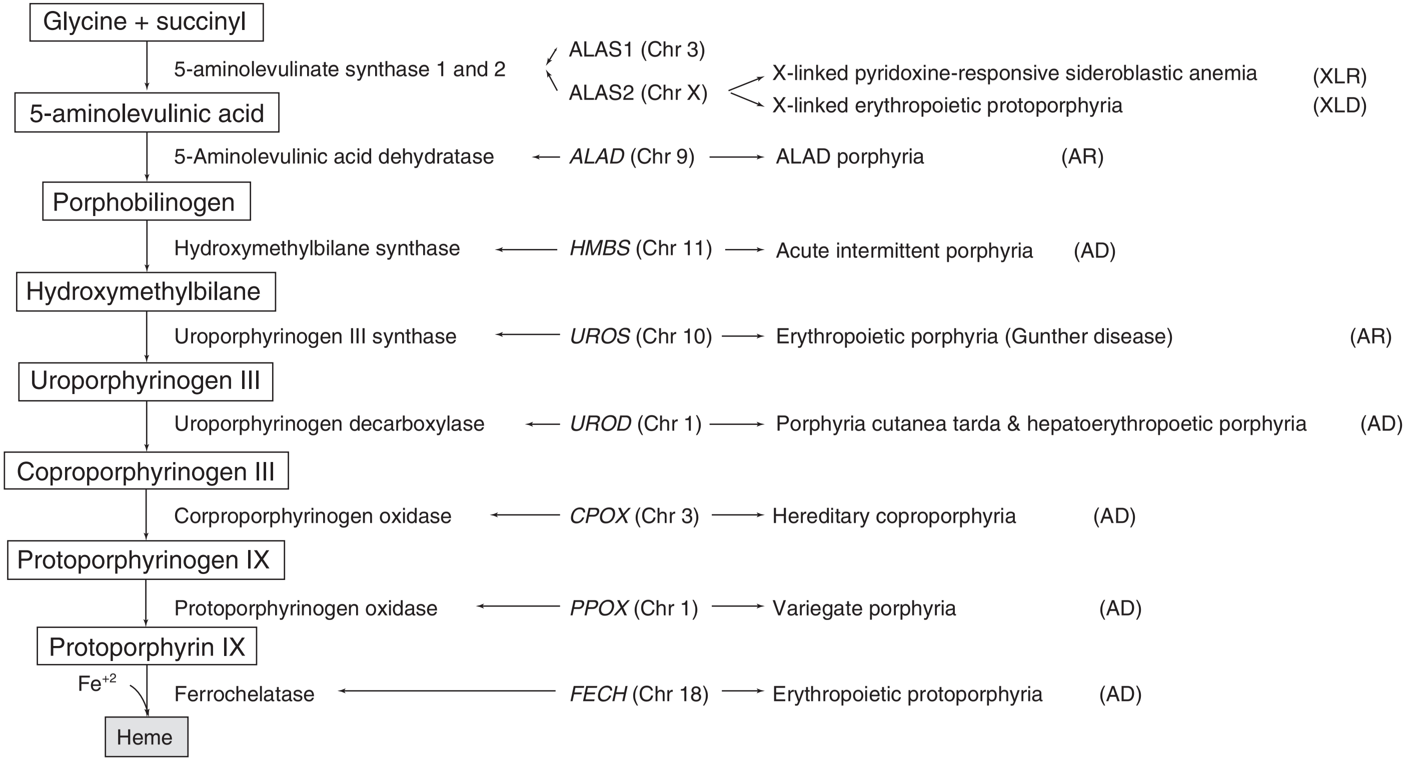

Glycine + succinyl
5-aminolevulinate synthase 1 and 2
ALAS1 (Chr 3)
ALAS2 (Chr X)
X-linked pyridoxine-responsive sideroblastic anemia (XLR)
X-linked erythropoietic protoporphyria (XLD)
5-aminolevulinic acid
5-Aminolevulinic acid dehydratase
ALAD (Chr 9)
ALAD porphyria (AR)
Porphobilinogen
Hydroxymethylbilane synthase
HMBS (Chr 11)
Acute intermittent porphyria (AD)
Hydroxymethylbilane
Uroporphyrinogen III synthase
UROS (Chr 10)
Erythropoietic porphyria (Gunther disease) (AR)
Uroporphyrinogen III
Uroporphyrinogen decarboxylase
UROD (Chr 1)
Porphyria cutanea tarda & hepatoerythropoetic porphyria (AD)
Coproporphyrinogen III
Corproporphyrinogen oxidase
CPOX (Chr 3)
Hereditary coproporphyria (AD)
Protoporphyrinogen IX
Protoporphyrinogen oxidase
PPOX (Chr 1)
Variegate porphyria (AD)
Protoporphyrin IX
Fe^{+2}
Ferrochelatase
FECH (Chr 18)
Erythropoietic protoporphyria (AD)
Heme

Fig. 4.7 Genes associated with forms of porphyria due to alterations in enzymes in the biosynthetic pathway of heme. The biosynthetic pathway of heme is presented on the left and highlights the enzyme that catalyzes each reaction. To the right the gene that encodes the respective enzyme and its chromosome mapping are shown. When pathogenic DNA variants have been reported in association with a form of porphyria, it is indicated as well as its inheritance pattern (AD: autosomal dominant; AR: autosomal recessive. XLD: X-linked dominant; XLR: X-linked recessive)

hepatoerythropoietic porphyria, and porphyria cutaneo tarda. Acute porphyrias include acute intermittent porphyria (AIP) and ALAD porphyria. They are characterized by short episodes of sudden onset affecting the central nervous system. Hereditary coproporphyria and variegate porphyria can present cutaneous and acute porphyria manifestations [25]. This clinical heterogeneity is associated with genetic heterogeneity as seen on the gene list of Fig. 4.7. The sensitivity to dosage of the enzymes in the heme pathway illustrates the phenotypic effects of distinct Mendelian inheritance patterns of pathogenic DNA variants leading to decreased heme biosynthesis and accumulated substrates (Fig. 4.7). Finally, the damaging products of the heme pathway may most often accumulate in the liver or skin, leading to the organ based classification of porphyrias.

AIP is the most common genetic type of porphyria, an autosomal dominant form caused by loss-of-function pathogenic variants in the *HMBS* gene, which codes for an enzyme with porphobilinogen deaminase function, known as hydroxymethylbilane synthase (Fig. 4.7). Clinically AIP is characterized by life-threatening, acute, severe, neuro-visceral abdominal pain without peritoneal signs, often accompanied by nausea, vomiting, tachycardia, and hypertension, that may be complicated by neurologic signs (mental confusion, seizures, peripheral neuropathy) and hyponatremia. Although often isolated, this clinical picture may be recurrent in up to 10% of the patients [25]. The genetic pathophysiology is explained by sensitivity to reduced dosage of the protein, a molecular definition of haploinsufficiency that should explain the other types of autosomal dominant porphyria (variants in *UROD*, *CPOX*, *PPOX* and *FECH* genes, Fig. 4.7). It is possible that enzymes encoded by the genes that cause autosomal recessive porphyria (*ALAD* and *UROS*) are not dosage-sensitive and their bi-allelic inactivation must occur to affect their function. It is also a possibility that complete loss of function of the *HMBS, UROD*, *CPOX* and *FECH* genes should be lethal in the embryo justifying the lack of autosomal recessive forms for these genes.

The first step in the heme biosynthetic pathway that occurs in mitochondria consists of condensation of glycine and succinyl COA in liver or erythroblasts. Two enzymes encoded by two distinct nuclear genes may catalyze this reaction. The *ALAS1* autosomal gene is expressed in the liver whereas the X-linked *ALAS2* gene is expressed in erythroblasts. No pathogenic DNA variant has been reported for *ALAS1*. However, pathogenic DNA variants in the paralog *ALAS2* have been associated with two phenotypes. Loss-of-function *ALAS2* variants cause X-linked pyridoxine-responsive sideroblastic anemia, whereas gain-of-function missense variants in *ALAS2* cause X-linked erythropoietic protoporphyria due to hyperactivation of the ALAS2 enzyme (Fig. 4.7) [25]. Although some authors avoid to classify X-linked erythropoietic protoporphyria as dominant inheritance due to asymptomatic heterozygous women, the clinical presentation is consistent with X-linked dominant inheritance. On one hand, the pathogenic variants in *ALAS2* are close to 100% penetrant in men. On the other hand, clinical heterogeneity and non-penetrant cases among women should be most probably due to X-inactivation effects (see Chap. 2), as fully penetrant cases are seen. Finally, gain-of-function effects are consistent with a dominant phenotype.

4.5 Genomic imprinting and Non-classic Mendelian Inheritance

Angelman syndrome (AS; OMIM 105830) and Prader Willi syndrome (PWS; OMIM 176270) are two clinically distinct neurodevelopmental disorders with opposite imprinting profiles, caused by reciprocal deletion of the human chromosome band interval 15q11-q13. Gene imprinting refers to the expression of a single allele of a diploid gene locus on a parent-of-origin-dependent basis due to regulated epigenetic silencing of the other allele (Box 4.2). The monoallelic expression of an imprinted gene can occur in all cells or be tissue-specific due to other layers of gene expression regulation, as previously discussed. AS is generally caused by loss of imprinted and maternally expressed genes in this region, specifically impacting the *UBE3A* (Ubiquitin-protein ligase 3A; OMIM 601623) gene. PWS can be due to deletion of 15q11-q13 paternal copy or imprinting control region (ICR), or maternal uniparental disomy of chromosome 15 (see Chap. 2). In addition, maternal CNV duplication at 15q11-q13 is also an important genetic predisposition to different neurodevelopmental disorders not associate with AS or PWS, featuring epilepsy, autism spectrum and psychotic disorders [12].

AS is characterized by severe developmental delay, intellectual disability, speech impairment, gait ataxia and/or limb tremor, and a unique behavior with an inappropriate happy demeanor that includes frequent laughing, smiling, and excitability. Microcephaly, seizures and decreased need for sleep are also common. Developmental delays are first noted at around age six months; however, the unique clinical features of AS do not become manifest until after age one year, and it can take longer before the clinical diagnosis is concluded.

PWS patients have reduced fetal activity, motor developmental delay, muscular hypotonia and feeding problems early in infancy. From the second year of life, they present extreme feeding problems characterized as hyperphagia, insatiable appetite and obsession with food, resulting in plethoric obesity. PWS is the most frequent genetic cause of life-threatening obesity. PWS patients may also present short stature, small hands and feet, hypogonadotropic hypogonadism, mild intellectual disability and behavior problems.

The ICR is a genomic regulatory region with few kilobases in length regulating epigenetic imprinting and consequently affecting the expression of target genes in an allele- or parent-of-origin-specific manner (Box 4.2). Its associated regulatory elements include differentially methylated regions and non-coding RNAs. Imprinted genes are commonly in clusters in a chromosome locus, constituting imprinted domains, controlled by an independent ICR [12].

After fertilization, there is a demethylation phase in which gamete DNA methylation patterns are erased on both parental genomes, except for certain genomic regions, including ICRs, that retain parent-specific DNA methylation. ICRs appear to have several binding sites for ZFP57 (zinc finger protein 57), which recruits the KAP1 (KRAB domain-associated protein 1) heterochromatic complex. After blastocyst implantation, *de novo* methylation takes place. Afterwards, the overall genomic levels of DNA methylation remain relatively stable except for primordial germ cells that undergo a second wave of genome-wide demethylation, including removal of parental epigenetic memory in ICRs. As germ cell development proceeds, cytosine methylation patterns re-establish at ICRs in specific oocyte or spermatozoon patterns, known as germline differentially methylated regions (gDMR), generally presenting histone marks associated with closed chromatin and gene repression. Maternally methylated gDMRs identified so far are more numerous than paternally methylated gDMRs, and tend to correspond to promoters, while the latter tend to function as insulators or enhancers. Different transcription factors recognize the gDMR methylated and unmethylated alleles directing differential epigenetic modification and imprinted expression of the locus. As consequence, in chromosomal imprinted domains, DNA methylation of specific genes controlled by the ICR differs between the maternally derived and paternally derived alleles, establishing somatic differentially methylated regions [12].

AS is caused by the lack of maternal *UBE3A* gene expression that includes *de novo* maternal deletions (~80% of cases), pathogenic DNA variants in the maternal allele (~10%), paternal uniparental disomy involving at least 15q11-q13 or chromosome 15 (up to 5%), and imprinting defects (up to 5%). *UBE3A* is only expressed from the maternally inherited allele in mature human neurons due to tissue-specific genomic imprinting by expression of *UBE3A* antisense transcript (*UBE3A-ATS*) from the paternally inherited allele, which silences the paternal allele of *UBE3A* in *cis* [12].

PWS paternal deletions (~60% of cases) encompass approximately 5–6 Mb at 15q11–q13, and maternal disomy of chromosome 15 account for 36% of cases. The remaining nearly 4% of PWS cases are due to paternal ICR defects owing to microdeletion or epigenetic alteration. PWS is caused by loss of the paternal expression of several contiguous genes at 15q11-q13. Maternal uniparental disomy of chromosome 15 and paternal ICR alterations are expected to double the expression of the maternally expressed gene *UBE3A*. In PWS patients, those two genotypes are more associated with psychotic illnesses than individual 15q11-q13 gene deletion genotypes [12]. As observed, AS and PWS show autosomal dominant inheritance although in most cases by a non-classic Mendelian pattern as their manifestations are generally not resultant of pathogenic DNA alterations but mostly dependent on epigenetic events.

4.6 Final Remarks: The Evolving Concept of Protein-Coding Genes

In the last 80 years, since the proposition in 1941 by the 1958 Nobel in Physiology or Medicine co-awardees Edward Tatum and George Beadle that one gene governs the production of one enzyme, later extended to the 'one gene—one polypeptide' hypothesis, Molecular Biology and, more recently, the field of Genomics have advanced tremendously. The 1990's genome projects as well as the large-scale genomic, transcriptome, and proteome studies of this millennium have broken down several paradigms. There are exceptions for every rule in Biology, making it challenging to define clear-cut concepts in the ever evolving field of Genomics. The chromosomal organization of the nearly 20,000 protein-coding genes of the human genome shows a number of complex loci with bidirectional promoters, sense-antisense gene pairs overlapping transcription in both DNA strands, some of which consisting of non-coding genes in the complementary strand. Moreover, recent high throughput transcriptome analyses have additionally identified transcripts initiating at enhancer *cis*-regulatory elements, still lacking a functional clarification. Thus the functional classification of promoters and enhancers have also become more controversial and a promising field for novel studies.

The protein-coding gene concept adopted here limits to the DNA length corresponding to the extension of the primary transcript. The baseline for this definition is that enhancers and silencers are short DNA elements that may modulate the transcription of protein-coding genes at distance, in variable locations that may be within other genes. Additionally, promoter elements are variable in sequence and promoter downstream motifs overlap the 5′ transcribed sequence of the gene. Likewise, promoter and enhancer chromatin signatures vary according to gene activity and consequently do not converge to a unifying classification associated with all protein-coding genes as reference. Thus, promoters, enhancers and silencers can be viewed as elements associated with the gene. The same reasoning applies to bidirectional promoters regulating two adjacent genes positioned in a head-to-head manner.

On the other hand, the adopted protein-coding gene concept challenges us in several frontlines. Alternative promoters may modify the length of the primary transcript, and disperse promoters as CpG islands may produce a few diffusely distributed TSSs. In its 3′ end, when the gene presents mutually exclusive exons containing alternative polyadenylation signals, alternative splicing confers the possibility of different primary transcripts for the same gene because their 3′ end will be co-transcriptionally, differentially trimmed.

Alternative splicing represents a large, if not the largest, source of transcriptome and proteome diversity allowing for more than one polypeptide isoform expressed by a gene. Genomes with more introns tend to have a greater variety of transcripts per gene. Recent studies have disclosed possible increase in proteome diversity independently on transcriptome variation due to translation of alternative ORFs on the same mRNA that harbors a single annotated CDS. This is an open research area very likely to contribute to the Genomics field in coming years, leading to the possibility of future discovery of an average number of translated ORFs per human protein-coding gene. Additionally, many cryptic exons deep in introns have either been annotated. If SNPs modify their splicing sites on a population basis, the in-frame expression of these cryptic exons may also contribute to increase diversity to the proteome. Moreover, non-polyadenylated circular RNAs have been described as an additional layer for regulation of gene expression. If translated, the resulting peptide may either have at least one exon-encoded amino acid sequence in differential N-to-C terminal orientation according to the CDS due to the RNA circularization or be in another reading frame. As they are uncapped, it is also expected that circRNAs should employ IRES for translation initiation.

Finally, as the ENCODE project catalyzes the annotation of protein-coding genes of the human genome, this task still appears endless as more global analyses of chromatin, transcriptome and proteome as well as functional assessment of genes, pseudogenes, associated elements, uORFs, and circRNAs will unveil their particularities, questioning established paradigms and challenging us to precisely define a protein-coding gene.

References

1. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A. 1961;47:1588–602.
2. Lander ES, Linton EM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
3. Dufner-Almeida LG, do Carmo RT, Masotti C, Haddad LA. Understanding human DNA variants affecting pre-mRNA splicing in the NGS era. Adv Genet. 2019;103:39–90.
4. Soutorina J. Transcription regulation by the mediator complex. Nat Rev Mol Cel Biol. 2018;19(4):262–74.
5. Suzuki Y, Tsunoda T, Sese J, et al. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. 2001;11(5):677–84.
6. Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance. Epigenetics. 2019;14(12):1141–63.
7. Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet. 2020;21:71–87.
8. Oldfield AJ, Henriques T, Kumar D, et al. NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. Nat Commun. 2019;10:3072.
9. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblast by defined factors. Cell. 2007;131:861–72.
10. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52(3):254–63.
11. Tolsma TO, Hansen JC. Post-translational modifications and chromatin dynamics. Essays Biochem. 2019;63:89–96.
12. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev Genet. 2019;20:235–48.
13. Cao A, Galanello R. Beta-thalassemia. Genet Med. 2020;12:61–76.
14. Moxley AH, Reisman D. Context is key: understanding the regulation, functional control, and activities of the p53 tumour suppressor. Cell Biochem Funct. 2020:1–13.
15. Khan D, Sharathchandra A, Ponnuswamy A, Grover R, Das S. Effect of a natural mutation in the 5′ untranslated region on the translational control of p53 mRNA. Oncogene. 2013;32: 4148–59.
16. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci. 1971;68:820–3.
17. Piovesan A, Antonaros F, Vitale L, et al. Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes. 2019;12:315–20.
18. Brunet MA, Levesque SA, Hunting DJ, Cohen AA, Roucou X. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. Genome Res. 2018;28:609–24.

19. Gehring NH, Roignant JY. Anything but ordinary–emerging splicing mechanisms in eukaryotic gene regulation. Trends in Genet. 2020;14
20. Singh RN, Ottensen EW, Singh NN. The first orally deliverable small molecule for the treatment of spinal muscular atrophy. Neurosci Insights. 2020;15:1–11.
21. The ENCODE Project Consortium, Snyder MP, Gingeras TR, Moore JE, et al. Perspectives on ENCODE. Nature. 2020;583:693–8.
22. Sisu C, Peia B, Lenga J, Frankishc A, Zhanga Y, et al. Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Si USA. 2014;111(37):13361–6.
23. Cheetham GSW, Faulkner J, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet. 2020;21:191–201.
24. Forslund SK, Kaduk M, ELL S. Evolution of protein domain architectures. In: Anisimova M, editor. Evolutionary genomics. Methods in Molecular Biology, vol. 1910. New York, NY: Humana; 2019.
25. Pischik E, Kauppinen R. An update of clinical management of acute intermittent porphyria. Appl Clin Genet. 2015;8:201–14.

Chapter 5
Noncoding Gene Families of the Human Genome

Ricardo Alberto Chiong Zevallos and Eduardo Moraes Reis

5.1 Noncoding Genes: Finding Treasure in Junk

The Human Genome Project (HGP) was an international effort aiming at the determination of the basic code, the genetic blueprint of the human being (see Chap. 1). Before the HGP, it was estimated that the human genome would have about 70,000–100,000 genes [1]. However, upon its completion the number of identified genes was only around 20,000 and it became clear that most of the human genome (>98%) do not encode proteins. Based on the general lack of evidence of strong purifying selection, the nonprotein coding fraction of the genome was initially dismissed as nonfunctional remnants of evolution, epitomized by the definition, "Junk DNA" [2]. However, it rapidly became clear the existence of regulatory DNA elements and noncoding RNAs being transcribed in intergenic regions and as time passed, the scientific quest to determine "what constitutes the human genome" shifted towards "what the genome does". Addressing this task required the establishment of new experimental approaches and computational techniques to model genome regulation, and in 2003 the public research consortium ENCODE (ENCyclopedia Of DNA Elements) was assembled to evaluate technologies to identify all functional elements present in the human and mouse genomes [3]. The two initial phases of the project established new methods and assigned biochemical function, as measured by RNA expression and the presence of regulatory chromatin marks, to more than 80% of the human genome in at least one cell type [4]. The ENCODE project is still ongoing and has just completed its phase 3, which generated RNA transcription, chromatin accessibility and modification, transcription

R. A. C. Zevallos · E. M. Reis (✉)
Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo,
São Paulo, SP, Brazil
e-mail: emreis@iq.usp.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_5

factor binding and DNA methylation datasets from hundreds of human and mouse cells lines, and annotated nearly a million of candidate cis-regulatory elements (cCREs) in the human genome that control the expression of at least 20,225 protein-coding and 37,595 noncoding genes [5]. While there is still an open debate concerning the extent of the non protein-coding portion of the genome that is transcribed by spurious RNA polymerase firing, and thus merely constitutes transcriptional noise [6], it is now clear that the previously called "junk DNA" comprises sequence elements that act as important signalers, mediators, and effectors in numerous mechanisms in normal physiological and pathological processes. In spite of the lack of protein-coding capacity, noncoding genes are transcribed producing noncoding RNAs (ncRNA) that can play essential roles in an organism's biology. The known functions of ncRNAs include gene expression regulation by recruitment or sequestering of transcription factors, the assembly of DNA-protein complexes, the degradation of mRNAs, transport of RNAs and even formation of cellular organelles. Also, the transcription levels of certain ncRNAs can be used as biomarkers for cancer detection or as a prognostic proxy to infer patient outcome or the response to a particular treatment, illustrating how these molecules can be explored to clinical ends. Only a small number of ncRNAs has been studied in detail and a deeper knowledge of their structural and functional properties is likely to reveal unanticipated molecular mechanisms that operate in human cells and that can be used to develop new applications in translational medicine.

5.2 RNA: A Versatile Molecule

In 1958, Francis Crick proposed the Central Dogma of Molecular Biology. In his model, the genetic information flows from DNA to RNA and from RNA to protein, but it could not go from proteins to nucleic acids. In the following decades, it was believed that proteins were responsible for all the enzymatic activities required to replicate the genetic content and the RNA was a mere intermediary of the genetic information between DNA transcription in the nucleus and translation of proteins in the cytoplasm. The discovery in the 80's of catalytic RNAs able to cleave transfer RNA precursors in *E. coli* [7] and the mechanism of exon self-splicing in ribosomal RNAs in *Tetrahymena* [8] shed new light to unexpected functions of RNA molecules and opened up a whole new perspective, regarding their role in the origin of life. The existence of RNAs with intrinsic enzymatic activity (i.e. ribozymes) able to process other RNA molecules implied that the first self-replicating systems did not necessarily have to be based on protein complexes after all. It provided support for a new theory to explain the origin of life replicative systems in which chemical evolution of polyribonucleotides eventually gave rise to ribozymes with RNA polymerase activity that were capable of self-replication. Thus, structural and catalytic RNAs found in modern organisms could be regarded as "molecular fossils" originated in a primordial world in which RNAs played a central *role* [9].

The functional versatility of RNAs is associated to their ability to acquire complex secondary and tertiary structures resulting from intramolecular canonical and noncanonical base pairing. Advances in experimental and computational methods allowed the structure determination of ribonucleic acid molecules providing new insights into their molecular mechanisms of action. The experimental approaches to investigate RNA's structure include biophysical methods such as X-ray crystallography, cryogenic electron microscopy and nuclear magnetic resonance spectroscopy [10], and computational methods for RNA secondary structure prediction based on folding minimum free energy or comparative sequence analysis, which can be significantly improved with the incorporation of spatial constraints obtained from experimental data, derived from structure-sensitive enzymatic cleavage and chemical probing reagents coupled to NGS [11, 12]. Post-transcriptional structural and chemical modifications have implications on RNA properties, such as stability, function and cellular transport/localization through interaction with different proteins. More than 150 RNA modifications have been described, which are under dynamic regulation by enzymes that add or remove specific modifications and proteins that target modified structures [13]. RNA modifications may elicit changes that affect the secondary structure, the ability to base pair with other nucleic acids or protein-RNA interactions [14]. Indeed, RNA modifications can alter the expression levels of the respective gene or even change the RNA function and in turn affect the expression of other genes along the genome in a downstream manner.

In eukaryotic organisms, RNA polymerase II transcribed RNAs are typically modified by 5′ capping, 3′ polyadenylation and splicing. The 5′ cap protects the RNA from degradation by 5′-3′ exonucleases and interacts with different protein factors involved in transport along the cell, splicing and mRNA translation. Also, the 5′ capping enhances the elongation of the nascent transcript by the interaction of a capping enzyme that allows the RNA polymerase to shift to a elongation-competent status [15]. Similar to the 5′ cap, addition of poly(A) tail to the 3′ end protects RNAs from degradation and is required to signal their export to the cytoplasm and efficient mRNA translation. Splicing, the removal of introns sequences and joining of the remaining exons, allows the generation of mature mRNAs encoding different polypeptides from identical precursors, due to alternative intron retention or exon skipping. All these modifications are critical for the function and diversity of protein-coding mRNAs and likewise must affect the availability and function of RNA Pol II transcribed ncRNAs. Pol II transcribed RNAs can undergo discrete single base modifications by RNA editing. Single base deamination can convert an adenosine into an inosine or a cytidine into an uridine, changing the final protein sequence encoded in mature mRNAs. Post-transcriptional editing may also decrease the ability of RNAs to form stable secondary structures or create/destroy microRNA binding sites at target molecules, which may affect the function or stability of protein-coding and ncRNAs alike. Modifications in RNA Pol I-transcribed transfer RNA (tRNA) molecules are the more diverse, involving a cascade of enzymes [14]. After transcription, tRNAs are subjected to removal of several consecutive bases, base substitutions, isomerization of uridines, base and ribose methylations, addition of sugars and complex organic adducts. The mature tRNAs may contain varying quantities of ring structures depending of which modifications were introduced, and may as well differ greatly in folding mechanism, stability and cellular localization. For example, single alterations in the anticodon loop, the hotspot of modifications in tRNAs, can modulate the fidelity of translation [14] . Modifications in ribosomal RNAs (rRNAs) are mostly changes of ψ torsion angle and 2′-O-methylations of the ribose [14]. Such modifications include more than a hundred of uridine isomerizations per molecule and in eukaryotes are assisted by small nucleolar RNAs (snoRNAs) that base-pair with the target rRNA and guide rRNA modification enzymes to determine the bases to be modified.

5.3 Noncoding Gene Families

High-resolution transcriptome sequencing studies have revealed thousands of novel ncRNAs with distinct properties and functions, highlighting the need to classify these molecules according to unifying features in order to facilitate their detailed study. While there is no consensus regarding the classification of ncRNAs [16], their (1) genomic organization, (2) their expression level variation across different tissues and (3) the transcript length are operationally useful properties largely used in the literature to group and annotate ncRNAs. The ENCODE project proposed a classification for ncRNAs based on their genome mapping coordinates that became widely accepted [17]. Noncoding RNAs are classified as "intergenic" if transcribed from a locus away from protein coding-gene locus or "intragenic" if transcribed from a region that overlaps the regulatory region or exon-intron structure of a different protein coding/ ncRNA gene (Fig. 5.1). Intragenic ncRNAs can be further subclassified based on the transcriptional orientation and overlap pattern with introns/

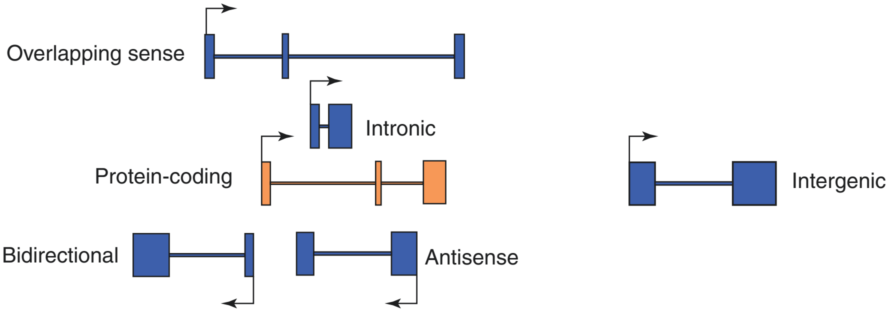


Fig. 5.1 Annotation of multi-exonic noncoding RNAs (boxes in blue) according to their genomic position relative to a protein-coding gene (in orange). Intragenic ncRNAs may present the same (overlapping sense ncRNAs, intronic ncRNAs) or opposite orientation (antisense ncRNAs) of the overlapping protein-coding gene. NcRNAs flanking protein-coding genes at close proximity may share regulatory regions being transcribed in the opposite direction from bidirectional promoters. Intergenic ncRNA loci do not overlap transcribed or regulatory DNA elements from protein-coding genes

exons of the host gene: antisense(as) ncRNAs are transcribed in the opposite direction of the host gene; sense overlapping ncRNAs are transcribed in the same orientation and contain an exon of the host gene spanning an intron; finally, intronic ncRNAs are transcribed in the same orientation within introns of the host gene.

The ncRNAs that are stably expressed across different tissues are referred to as "housekeeping" RNAs, to distinguish them from transcripts with expression restricted to specific tissues or that change during developmental, physiological or pathological states. The latter include ncRNA classes that interact with the cellular machinery to modulate gene expression and thus are referred to as "regulatory" ncRNAs. Finally, ncRNAs are generally classified as short ncRNAs (sncRNAs) if their primary sequence contains up to 200 nucleotides, or long ncRNAs (lncRNAs), if containing more than 200 nucleotides in length. These features will be used to organize the description of the known ncRNAs gene families in the following sections, which are depicted in Table 5.1and in their cellular context in Fig. 5.2.

Table 5.1 Main noncoding RNA classes in the human genome

	Size (base)	Genomic context	Number of loci	Functions	References
Housekeeping RNAs					
tRNAs	<0.1 kb	Clusters, mainly in chromosomes 1 and 6	>500	Codon translation	[18–20]
rRNAs	121–5070 nt	Tandem arrays units of gene repeats encoding rrna contained in nucleolar organizer regions on the five acrocentric chromosomes	>400 units tandemly arrayed	Ribosome assembly, protein synthesis	[4, 21, 22]
snoRNAs	60–300 nt	Mainly in introns of other genes	>300	rRNA modification	[23]
snRNAs	100–200 nt	Widespread loci	>1800	mRNA splicing, transcriptional elongation	[24, 25]
Telomerase RNA	451 nt	Chr 3	Single copy	Template-assisted reverse transcription of chromosome ends by telomerase	[26, 27]
SRP RNA	299–302 nt	Chr 14	3	Component of the signal recognition particle RNA-protein complex, which targets transmembrane and secretory proteins to the endoplasmic reticulum	[28]

(continued)

Table 5.1 (continued)

	Size (base)	Genomic context	Number of loci	Functions	References
Small ncRNAs					
piRNAs	26–31 nt	Clusters, intragenic loci	23,439	Mobile elements repression	[29]
miRNAs	19–24 nt	Inter- or intragenic loci	>1424	Post-trancriptional gene silencing	[29]
tiRNAs	17–18 nt	Downstream of tsss	>5000	Regulation of transcription initiation	[30]
PASRs	22–200 nt	5′ regions of protein-coding genes	>10,000	Unknown	[31]
TSSa-RNAs	20–90 nt	–250 and +50 bp of tsss	>10,000	Maintenance of transcription	[32, 33]
PROMPTs	<200 nt	−205 bp and − 5 kb of tsss	Unknown	Initiation of transcription, enforces promoter directionality	[34–36]
Long ncRNAs					
lincRNAs	>200 nt	Widespread loci	>1000	Recruiting of chromatin modifying factors to specific locus, miRNA sponge, RNP assembly	[37–40]
T-UCRs	>200 nt	Widespread loci	>350	Subset of highly conserved lncRNAs. Transcriptional regulation as miRNA sponges.	[41]

5.3.1 Housekeeping ncRNAs

Housekeeping RNAs are constitutively transcribed and accumulate at high levels in virtually all cells of an organism and are essential for vital functions [42]. In eukaryotes, these include ncRNAs involved in the translational apparatus (tRNAs, rRNAs), and the subcellular localization of newly synthesized proteins (SRP RNA), and nuclear localized RNAs required for transcriptional control and post-transcriptional processing/modification of other RNAs (snRNAs, snoRNAs).

Mature transfer RNA (tRNAs) molecules are clover-leaf-shaped and have approximately 80 nucleotides [2]. Through covalently attachment to its 3′ end, the tRNA carries a specific amino acid to a ribosome aiming the incorporation into a growing polypeptide chain. The tRNA-amino acid specificity depends on the base-pairing of its anticodon region with the triplet codon of the mRNA being translated. Specific aminoacyl tRNA synthetases catalyze the esterification between the 3′-OH of tRNA and the corresponding amino acid in the cytoplasm. There is only one aminoacyl tRNA synthetase for each tRNA, but some tRNAs can base-pair with more than one codon through wobbling base-pairing. Due to the genetic code degeneracy, some triplet codons code for the same amino acid and 49 distinct tRNAs

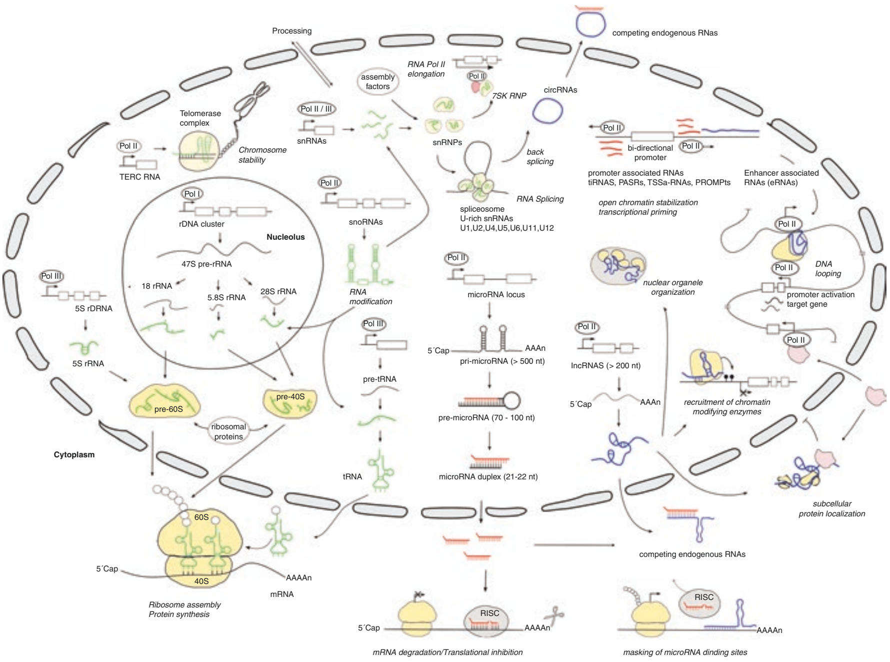


Fig. 5.2 Illustrative image of the main ncRNAs classes expressed in human cells and their structural and regulatory molecular functions: Housekeeping ncRNAs (green), small ncRNAs (red), long ncRNAs (blue)

transcribed by nearly 500 nuclear genes (plus 22 distinct tRNA genes in the mitochondrial genome) are sufficient to read the 61 codons for the 21 proteinogenic amino acids (20 of the standard genetic code, plus selenocysteine).

Ribosomal RNA (rRNA) are structural and catalytic components of the two-subunit ribonucleoproteic complexes that form the ribosome. In humans, the 40S small subunit is constituted by 33 proteins and the 1900 nt-18S rRNA. The 60S large subunit contains 49 proteins and 3 rRNAs with 120 nt (5S), 160 nt (5.8S) and 4700 nt (28S), respectively. Together the two subunits form the a 80S ribosome [2]. The 28S rRNA is a ribozyme with peptidyl transferase activity that catalyzes the formation of the peptide bond between the incoming amino acid and the polypeptide chain being synthesized in the ribosome. Ribosomal RNA genes are present in multiple copies in human genome (˜400), comprising tandemly repeated units localized in nucleolar organizing regions (nucleolus), where ribosome biogenesis and maturation take place. Each unit encodes a 13-kb precursor 45S rRNA that is processed to generate the 28, 5.8, and 18S rRNA, while 5S rRNAs are transcribed from separate loci [21]. Mature functional rRNAs are edited during processing and display 115 methyl group and 95 pseudouridine modifications. Mitochondrial (70S) ribosomes contain two rRNAs (12S and 16S) and are more similar to the ones found in bacteria.

Secretory and membrane proteins are targeted to the protein translocation apparatus of the cell by a signal recognition particle (SRP) ribonucleoprotein complex that associates with leading signal peptides and with specific receptors in the endoplasmic reticulum membrane. The SRP complex is evolutionarily conserved and in mammalian cells is comprised of 6 proteins and a ~300 nt 7S RNA that act as a scaffold and is critical for SRP complex assembly and function [43].

Several nuclear localized, stable small ncRNAs ranging from 60 to 300 nt in length accumulate in the nucleus and are generally denominated small nuclear RNAs (snRNAs). These include 9 highly conserved uridine-rich spliceosomal snRNAs that associate with proteins to form the major (U1, U2, U4, U5, U6) and minor (U11, U12, U4atac, U6atac) snRNP spliceosomal complexes that catalyze the removal of intronic sequences during constitutive and alternative pre-mRNA splicing [44]. A subtype of nucleolar localized snRNAs, small nucleolar RNAs (snoRNAs) associate with proteins to form snoRNPs whose primary function is to drive modifications in pre-ribosomal RNAs (pre-rRNAs) by RNA-RNA base-pairing and recruitment of RNA modifying enzymes to target sites in rRNAs, whereas some snoRNAs may also cleave pre-rRNAs and participate in the modifications of small nuclear RNAs and mRNAs [2]. There are two classes of snoRNAs: C/D box snoRNAs participate in methylation and the H/ACA box snoRNAs guide pseudouridylation of pre-rRNAs. SnoRNAs are transcribed from hundreds of distinct loci, frequently located in intronic regions and being co-expressed with housekeeping genes involved in ribosome biogenesis or function, which suggests a common evolutionary process [45]. Small Cajal body-associated RNAs (scaRNAs) have similar biogenesis and guide the modification of spliceosomal snRNAs in nuclear Cajal bodies [46]. The 7SK RNA (~330 nt) is an abundant, evolutionarily conserved snRNA that interact with and modify the activity of the positive transcription elongation factor b (P-TEFb) to favor RNA pol II elongation and gene expression [47].

5.3.2 Small Noncoding RNAs

In 1988, Andrew Fire and Craig Mello received a Nobel prize for revealing the mechanism of RNA interference (RNAi) in which double-stranded RNA (dsRNA) molecules formed by complementary sense and antisense RNA stretches of a given gene, when introduced in the worm *C. elegans* were able to silence specifically the corresponding mRNA [48]. The cellular machinery involved in the RNAi has been extensively studied in the following years in various organisms and it involves the generation of small interfering RNAs (siRNAs) in the range of 20–25 nucleotides from longer dsRNA precursors. The RNAi mechanisms act at both transcriptional and posttranscriptional levels and its effectors endogenously generated or introduced in the cell by infecting viruses. In mammalian cells RNAi plays fundamentals roles in the maintenance of cellular homeostasis by assuring genome stability and the fine tuning of gene expression levels in response to physiological changes. The

RNAi mechanisms act at both transcriptional and posttranscriptional levels and its effectors in physiological processes in mammalian cells include piwi-interacting RNAs (piRNAs) and microRNAs (miRNAs) [42].

The piRNAs are 26–31 bp transcripts with expression restricted to the testis and associate with Piwi proteins, which are essential for spermatogenesis in germline cells. The piRNA-Piwi protein complex induce silencing of DNA mobile elements (transposons) preventing them from causing genome instability acting both at the transcriptional level by inducing the formation of heterochromatin in the nucleus and post-transcriptionally by promoting the degradation of transposon-derived RNAs in the cytoplasm. PiRNAs originate from RNA precursors transcribed from gene clusters or independent loci that contain repeated transposon sequences, and when associated to Piwi-proteins are able to guide the silencing complex to target sites through its ability to form specific base-pairing with nascent mature transposon-derived RNAs. Aberrant expression of piRNAs has been detected in tumors suggesting that it may play a role in the genomic instability that is a hallmark of the disease [49].

The mature miRNAs are approximately 22 nucleotide-long molecules that guide a post-transcriptional gene expression regulatory mechanism by base-pairing with target mRNAs to inhibit translation and destabilize the RNA molecule. The molecular mechanisms and biological functions of miRNAs in developmental and pathological processes are well established. For review see [50, 51]. The miRNA biogenesis initiates with transcription by RNA Pol II of long primary transcripts (pri-miRNA). Pri-miRNAs can be thousands of nucleotides long, may have promoter and enhancer elements similar to protein-coding genes (see Chap. 4) and may produce multiple miRNAs. Pri-miRNAs are subsequently processed in the nucleus by the Drosha ribonuclease, generating 70–100 nt precursors (pre-miRNA) that are able to assume a hairpin secondary structure. Alternatively, pre-miRNAs can also be located in introns and be co-transcribed with the host gene, being excised during splicing. In animals, pre-miRNA is transported to the cytoplasm, where it will be processed by the ribonuclease Dicer to generate a 19–24 nt long double-stranded mature miRNA. One of the strands (guide strand) is loaded onto the RNA-Induced Silencing Complex (RISC). The selection of the guide strand is dictated by the thermodynamic stabilities of the two duplex ends, being favored for incorporation into the RISC the strand having its 5′ terminus at the less stably base-paired end of the duplex. The activated RISC is guided by base pair complementarity of the miRNA and binding sites in target RNAs, often located in 3′-untranslated region of mRNAs in animal cells, to elicit their translational repression or deadenylation and degradation. Less often in mammalian cells but frequent in nematodes and plants, the perfect base-paring between the miRNA and target mRNA activates the endonucleolytic cleavage of the target by RISC Argonaute protein. A stretch of complementary bases of only 6 nucleotides (seed sequence) with the target is sufficient to activate RNAi and therefore a single miRNA can often target multiple different mRNAs. Also, a target mRNA may have multiple binding sites for more than one miRNA and the effective down regulation of the target may require the cooperative effect resulting from RISC binding to adjacent biding sites [52]

Different types of small ncRNAs that map to the vicinity of the transcription start sites (TSS) of actively transcribed genes have been reported in eukaryotic cells. These include Promoter-Associated Short RNAs (PASRs), Transcription Start Site-Associated RNAs (TSSa-RNAs), and Transcription Initiation RNAs (tiRNAs) (reviewed in [33]). It has been proposed that promoter-associated sncRNAs are generated by RNA Pol II pausing before productive elongation takes place [33] but it is unclear if all shares a common biogenesis pathway since there are differences in length (ranging from 18 to over 200 nt), mapping position (−250 to +50 to TSS) and presence/absence of 5′cap. Unstable longer transcripts originated upstream (0.5 to 2 kb) of TSSs in both strands of cells with defective exosome function can be a source of TSS-aRNAs [33], and thus are functionally related to these promoter-associated sncRNAs. In common they all appear to be unstable RNA Pol II transcripts that are generated from bidirectional promoter activity and to contribute positively for gene transcription. It has been documented that most human promoters drive transcription in both directions to produce small RNA transcripts, but transcriptional elongation proceeds only in the direction of the annotated downstream gene [32]. Also, the abundance of promoter-associated sncRNAs correlate with the gene expression level and with RNA Pol II occupancy at the TSSs of the corresponding loci [34]. Mechanistically, one study provided evidence that promoter-associated sncRNAs generated during RNA Pol II pausing are kept bound to the complex and may contribute to maintain/regulate local gene expression through the recruitment of transcriptional regulators that are required for transition into productive elongation [53]. Other studies showed that promoter-associated sncRNAs may regulate transcription by modulating the epigenetic landscape by affecting the methylation status of CpG-rich promoter regions [54] or the accessibility of chromatin modifiers [55] of target genes. PROMPTs, a class of promoter-associated ncRNA, has been reported as enriched in CpG-enriched promoter regions and capable of modulating the local DNA methylation density [54].Taft and colleagues first noted that genomic binding sites of the chromatin regulator CCCTC-binding factor (CTCF) colocalize with RNAPII and are highly enriched for tiRNAs [55]. Depending on the chromatin context, CTCF may work as a transcriptional activator, a repressor or as an insulator protein, physically blocking communication between enhancers and gene promoters. Next, it was observed in breast tumor cells that depletion of tiRNAs overlapping a CTCF binding site located in the *CDKN1* locus promoted the recruitment of CTCF, increased nucleosomal positioning at the locus and increased expression of the encoded p21 protein, possibly through the deposition of H3K4me3 transcription activation marks in the CTCF-adjacent nucleosomes [55].

5.3.3 Long Noncoding RNAs

The high-resolution analysis of the transcriptome of eukaryotic cells have revealed the existence of thousands of RNAs with hundreds to thousands of nucleotides in length that do not encode proteins. As mentioned above, long noncoding (lnc)RNAs are classified according to their localization relative to protein-coding genes. Large intergenic non-coding RNAs (lincRNAs) are transcribed from regions between genes whereas intragenic lncRNAs overlap with exons, introns or 5′- and 3′-untranslated regions of protein-coding gene loci. LincRNAs show a higher conservation level than generic repeat sequences, meaning they are not evolutionarily neutral, but their rapid evolution rate and low interspecies conservation make it difficult to identify them by ab initio sequence analysis methods alone [56]. LncRNAs comprise a heterogeneous class of molecules not yet fully catalogued. Unlike small ncRNA classes that have already been extensively studied and whose biogenesis and mechanisms of action are already well known (e.g. microRNAs and piRNAs), there are relatively fewer lncRNAs that have been characterized in detail. However, it is evident that lncRNAs play important regulatory roles in the control of gene expression at the transcriptional and post-transcriptional levels, and participate in central biological processes of multicellular organisms, such as gene dose compensation, cell differentiation and organogenesis [57]. Among the emerging mechanisms of action of lncRNAs, it is highlighted (1) the ability to act as a decoy by sequestering regulatory proteins that bind to DNA, (2) the recruitment of proteins to form ribonucleoprotein complexes capable of interacting with chromatin or serving as scaffold for the organization of subcellular supra-molecular complexes, (3) acting as guides, directing protein complexes to specific sites in the genome or (4) acting as endogenous competitors for the binding of microRNAs [58, 59]. The 6200 nt lncRNA HOTAIR (Hox Transcript Antisense Intergenic RNA) illustrates an archetypical mechanism by which lncRNAs regulate gene expression in trans by recruiting chromatin modifying complexes to specific loci distinct from its site of transcription. The Homeobox (Hox) loci comprises several genes encoding homeotic transcription factors that play key roles in controlling the body plan along anterior–posterior axis and their transcription is tightly controlled during embryonic development. Expression of HOTAIR from the *HOXC* locus is temporally regulated and required for the recruitment of the Polycomb repressive complex 2 (PRC2) that catalyzes the trimethylation of histone H3 at lysine 27 (H3K27me3) to establish a repressive chromatin state across 40 kb of the adjacent *HOXD* locus [60]. PRC2 can interact with many different RNAs [61] and it is conceivable that other lncRNAs in addition to HOTAIR will be revealed in the future that regulate gene expression in trans by the recruitment of PRC2 to distinct genomic loci. In contrast to HOTAIR, HOTTIP (HOXA distal transcript antisense RNA) is an intergenic lncRNA (˜3700 nt) transcribed from the 5′ end of the *HOXA* locus that activates in *cis* gene expression of multiple HOXA genes [62]. From its transcription site, HOTTIP interacts and recruits the adaptor protein WDR5 of the mixed lineage leukemia (MLL) complex through chromosomal looping, that catalyzes histone H3 lysine 4 trimethylation (H3K4me3), an activation epigenetic mark, across the *HOXA* locus [63]. Interestingly, the HOTTIP locus contain both H3K4me3 and H3K27me3 modifications, bivalent pattern associated with poised regulatory sequence. Therefore, HOTTIP interaction with WDR5-MLL complex constitutes a positive feedback loop and H3K4me3 enrichment is associated with *HOTTIP* transcription. HOTTIP also interacts with PRC2, which catalyzes H3K27me3, possibly constituting a negative feedback loop since decreased H3K27me3 at HOTTIP is associated with HOTTIP transcription [64]. The MLL protein family is responsible for regulating expression patterns of several Hox genes, contributing significantly to many cell fate commitments during development and diseases establishment. Moreover, genetic rearrangements of MLL genes and post-translational regulation, such as recruiting by lncRNAs, are associated with several aggressive human leukemias [65].

A lincRNA that has been extensively characterized is Xist (X-inactive specific transcript), a 17 kb lncRNA on the X chromosome that is critical for establishment of epigenetic silencing of the X chromosome copies in female mammals as a mechanism of dosage compensation (see Chap. 2) [66]. In the X chromosome that will be silenced, transcription of Xist RNA promotes the recruitment in *cis* of Polycomb Repressor Component 2 (PRC2) through the interaction of a repeated element on its 5′ end (repeat A) with the histone-lysine N-methyltransferase EZH2, the catalytic subunit of PRC2. Xist RNA spread along the chromosome drives the PRC2-mediated deposition of repressive HeK27me3 marks in nucleosomes and initiates chromatin changes that result in heterochromatin formation and X-inactivation [66]. Interestingly, an RNAi-based mechanism involving an lncRNA antisense to Xist, Tsix (37 kb), also contributes for X-inactivation [67]. In the active X chromosome, Tsix inhibits EZH2-Xist binding. Ogawa and co-workers proposed a model in which Xist and Tsix anneal, forming a RNA duplex processed by Dicer resulting in small RNAs required to repress the chromatin inactivation of the X chromosome that remains active. Prior to inactivation, both X chromosomes would generate Tsix:Xist duplexes and the constant Tsix expression of one X would result in their Dicer-dependent processing. The small RNAs would act in *cis* silencing Xist in the active X chromosome preventing its inactivation. The different, but plausible, models of XCI suggest a complex RNA network regulating chromatin modifications. It is worth mentioning that several lncRNAs contain functional repeat sequence domains [68].

The lncRNA Airn (for Antisense Igf2r RNA Noncoding) promotes genomic imprinting of proximal and distal genes in trans by mechanisms of chromatin-mediated repression, as described above, but also in *cis* through transcription interference (TI). In TI, the elongating DNA-dependent RNA polymerase (RNAP) from an upstream transcriptional unit elongates its product normally until it interacts with molecular components of a downstream transcriptional process, such as another RNAP, structural chromatin proteins or transcription factors bound to DNA elements. Importantly, the dynamics of chromatin modifications may be pivotal for promoting transcriptional interference by modifying the spatial reorganization and positioning the transcriptional machinery from different loci in closer proximity [64]. Airn is expressed only from the paternal allele and its expression silences the maternal Igf2r (insulin-like growth factor 2 receptor) gene cluster by two mechanisms according to proximity. Airn inhibits proximal overlapping genes by TI [69] and distal non-overlapped genes by recruiting repressive chromatin modifying complexes PRC2 and EHMT2 [70]. Human Igf2r has imprinted expression in fetal tissues, but not in adult tissues, highlighting the importance of the Airn-mediated gene expression suppression in developmental processes [71].

Long noncoding RNAs also play structural roles in the organization and function of large nuclear ribonucleoprotein complexes that participate in the processing and metabolism of other RNAs. Nuclear speckles, also called interchromatin granule clusters, are supramolecular structures involved in splicing factor storage and modification in which the ~8700 nt lncRNA MALAT-1 is required for assembly of speckle proteins and directly regulates the phosphorylation of splicing factors affecting their function [72] . Likewise, paraspeckles are ribonucleoprotein complexes present in the nucleoplasm in the form of discrete foci whose biogenesis and maintenance depends on the lncRNA NEAT1 (MEN ε/β) [73]. Although the cellular functions of paraspeckles have not yet been fully characterized, these structures are involved in the reversible nuclear retention of RNAs and in the flow of messenger RNAs into the cytoplasm [74].

Intragenic ncRNAs comprise many different subclasses of transcripts, some of them with overlapping structural features and properties. Antisense(as) RNAs are transcribed from the opposite strand of a host protein-coding or noncoding loci and overlap with exonic or regions of the host gene, eventually spanning intron/exon boundaries. Between 50% and 70% of the protein-coding gene loci host independently transcribed asRNAs. The complementary nucleotide sequence of asRNAs allow them to base-pair with the sense mRNA and offer an opportunity for *cis*-regulation by transcriptional interference [75]. Intragenic lncRNAs transcribed with the same strand orientation of the mRNA may span intron/exon boundaries (sense partially intronic RNAs or PINs) or be totally located within intron boundaries (totally intronic RNAs or TINs) [76]. Of note, TINs represent almost 70% of the nuclear encoded ncRNAs (non-rRNA) and 40–50% of all cellular (non-rRNA) RNAs mass [76].

Some lncRNAs act in *cis*, near the locus from which they are transcribed, while others act in trans, at remote regions of the genome, virtually with no distance limitation [12]. As described in Chap. 4, enhancers are *cis*-regulatory DNA elements that can interact through chromosome looping with promoters located thousands of bases away and elements of the transcription machinery. Recent findings revealed that some enhancers generate lncRNAs (enhancer-associated ncRNAs, enhancer RNAs or enhancer-derived RNAs) that contribute to the enhancer function and it has been proposed that the act of transcription in enhancer regions contributes to stablish an open chromatin state and formation of chromosome looping that facilitates transcription [77]. As an example of functional eRNAs, Li et al. reported that 17b-oestradiol (E2) binding to estrogen receptor α (ERα) in a breast tumor cell line correlated with the increased transcription of eRNAs from enhancer elements near coding genes upregulated by E2. In the proposed model, eRNA transcription increases after signal-dependent activation of the local enhancer and the eRNA facilitates enhancer-promoter looping by stabilizing the interaction of the Mediator complex and cohesin, that in turn facilitate loading of RNA Pol II at the target gene's promoter [78]. The knockdown of eRNAs transcribed from ERα-related enhancers at *NRIP1* (Nuclear receptor-interacting protein 1) or *GREB1* (Growth regulating estrogen receptor binding 1) loci decreased the interaction between enhancers and promoters and the downstream activation of these genes [78]. It should be noted that both eRNAs and lincRNAs loci are localized in intergenic regions and the classification of an actively transcribed noncoding locus can be ambiguous. The majority of eRNAs are unspliced, non-polyadenylated bidirectional transcripts (2D-eRNA) but there are cases of polyadenylated and unidirectional eRNAs (1D-eRNA). In general, eRNAs are transcribed from enhancer regions defined by enrichment of H3K4me1 and low H3K4me3 content [77]. Conversely, lincRNAs are usually polyadenylated and spliced, transcribed unidirectionally from promoters rich in H3K4me3 and with low H3K4me1 content. To illustrate the difficulty in distinguishing eRNAs from lincRNAs, ncRNA-a (ncRNAs activating), a set of long (˜800 nt) noncoding RNAs enriched in H3K4me3 and H3K36me3 marks, have been initially annotated by GENCODE as lincRNAs and later showed enhancer-like properties [77].

Circular RNAs (circRNAs) constitute a recent and still poorly characterized class of noncoding RNAs that result from splicing of primary RNA precursors in reverse order (backsplicing) producing covalently closed loop structures without neither 5′–3′ polarities nor terminal modifications found in linear RNAs such as 5′ cap and 3′ poly-A-tails [79]. Most circRNAs arise from an upstream exon 3′ splicing donor site joining a downstream exon 5′ acceptor site (ecircRNAs), but it has been reported the existence of circRNAs originated by debranching of intron lariats followed by circularization (ciRNAs) [80]. The circular nature of cirRNAs confers increased stability since they are resistant to exonucleolytic degradation. The functions attributed to these molecules rely on the fact that they retain functional elements present in the host mRNA. Thus, circRNAs containing binding sites for splicing factors or translation start sites may regulate the alternative splicing or the translation rate of the host mRNA by competing for splicing factors or ribosome binding. Likewise, circRNAs may retain binding sites for miRNAs present in the host mRNA and thus, compete for miRNA binding, acting as a “miRNA sponge” or competing endogenous RNA (ceRNA). This later function is not restricted to circRNAs and is also attributed to other lncRNAs, pseudogenes (see Chap. 4) and even protein-coding mRNAs [81], which would result in highly interconnected transcriptional networks comprised by miRNAs, target mRNAs and various types of ceRNAs sharing the same miRNA binding sites. Such ceRNA-based networks add an addition layer of complexity for gene expression regulation and its significance for maintenance of cellular homeostasis and the pathophysiology of human diseases is currently under intense investigation [82].

5.4 Clinical Relevance of ncRNAs

Compared to protein-coding RNAs, lncRNAs have a greater tissue specificity, highlighting its potential as biomarkers in several diseases including cancer [83]. Remarkably, although several small and long noncoding RNAs have been associated to cellular phenotypes that support tumor growth and progression (Table 5.2), there are only few examples that have already translated into the clinical practice. A successful example of non-invasive diagnostic application of a lncRNA is the

Table 5.2 Examples of small and long noncoding RNAs with oncogenic or tumor suppressor roles in cancer hallmarks

Cancer hallmark	Small ncRNAs	Long ncRNAs
Sustaining proliferative signaling	miR-17 ~ 92 suppress PTEN [84, 85] micro RNA let-7 suppress Ras [86, 87] snRNA RN7SK regulates gene transcription elongation by binding to the positive transcription elongation factor b (P-TEFb) and masking its cyclin-dependent kinase-9 activity [88]	SRA is a coactivator for the steroid receptors PR, ER, GR and AR [89] PCAT-1 regulates cell proliferation, apoptosis, migration and invasion, serves as scaffold to Polycomb Repressor Complex 2 (PRC2) and guide it in trans [90] ncRNAs derived from cell cycle gene promoters [314] KRAS1P act as miRNA sponge [91] PR antisense regulates gene expression [92]
Evading growth suppressors	miR-675 inhibits pRB, interfering on the cell cycle arrest [93]	PSF-interacting RNA binds to protein PSF, releasing PSF from a oncogene, which is activated [94] ANRIL recruits PRC1 and PRC2 [95, 96] GAS5 induces cell arrest and apoptosis [97] lincRNA-p21 represses p53 targets, inducing apoptosis [98] E2F4 antisense represses protein levels of the E2F4 cell cycle repressor [99]
Enabling replicative immortality	miR-34a promotes senescence in colon cancer cells [100]	TERC the RNA component of the telomerase is amplified in tumors [101] TERRA (telomeric repeat-containing RNA) acts as negative regulator of telomerase [315]
Activating invasion and metastasis	miR-10b promotes migration, invasion, and metastasis [102] miR-200 inhibits EMT [103, 104]	HOTAIR may act as ceRNA to positively regulate HER2 in gastric cancer [105]; associates and retargets PRC2 for epigenetic silencing of metastasis suppressor genes [106] HULC acts as miRNA sponge [107]
Inducing angiogenesis	miR-296 inhibits HGS in tumor associated endothelial cells, enhancing angiogenesis [108]	HULC promotes tumor angiogenesis in liver cancer by up-regulating sphingosine kinase 1 [109]
Resisting cell death	miR-21 inhibits PDCD4, suppressing caspase activation [110]	PCGEM1 inhibits apoptosis [111] CUDR confers resistance to doxorubicin and etoposide [112] uc.73A(P) induces drug resistance through inhibiting apoptosis [113] SPRY4-IT1 inhibits apoptosis and promotes cell proliferation and invasion [114] PANDA interacts with transcription factor NF-YA, reducing expression of pro-apoptotic genes [115]

(continued)

Table 5.2 (continued)

Cancer hallmark	Small ncRNAs	Long ncRNAs
Genome instability and mutation	Deregulation of several microRNAs promotes genome instability by affecting cell cycle, mitosis and DNA damage repair [116] instability,baffecting cell cycle, mitosis and DNA damage repair [116]	NORAD null mutants develop genomic instability and aneuploidy by hyperactivation of PUMILIO proteins [117]

urinary PCA3 test approved by the US Food and Drug Administration (FDA) for prostate cancer diagnosis. The intronic lncRNA PCA3 (prostate cancer antigen 3) is highly expressed in prostate cancer and can be detected in urine samples by quantitative RT-PCR, along with PSA transcript levels, which has an equivalent expression pattern in tumor and benign cells [118]. The ratio of PCA3 to PSA transcripts is used to calculate a prognostic score, but its sensitivity and specificity is still controversial. The PCA3 test has been used to assess the need of re-biopsy in patients who kept elevated PSA levels after a negative biopsy [119]. Since PSA, but not PCA3, levels are influenced by other clinical conditions, the PCA3 test is also useful in prostate cancer cases with PSA overexpression and suspicion of chronic prostatitis [118]. Several other lncRNAs have been shown to bear prognostic potential in different types of cancer. Zhan et al. [120] investigated exosome-derived lncRNAs isolated from patient urine samples and identified lncRNA PCAT-1 as associated to recurrence-free survival of non-muscle-invasive bladder cancer [120]. In agreement with this result, Cui et al. analyzed tissue samples from gastric cancer patients and discovered that high levels of PCAT-1 were associated to poor overall survival, suggesting it could be a valuable prognostic biomarker for gastric cancer [121]. Similarly, Zhao et al. found that PCAT-1 increased expression in endometrial carcinomas was positively correlated with limited patient survival [122]. Another example of lncRNA with promising results for cancer diagnosis is the lncRNA LINC01535. In a recent study, Song et al. [123] showed that LINC01535 promotes cervical cancer progression by disrupting a miR-214/EZH2 double negative regulatory loop [123]. High LINC01535 expression levels are correlated with advanced-stage poor prognosis of cervical cancer. LINC01535 acts as ceRNA binding to miR-214 and releasing the repression of EZH2, with an oncogenic effect that promotes cell growth, migration and invasion *in vitro* and xenograft growth *in vivo*. The tumor suppressor miR-214 was reported to be downregulated in several cancers, such as oesophageal squamous cell carcinoma, papillary thyroid carcinoma, breast and colorectal cancer [124]. Thus, LINC01535 is a competitive endogenous RNA with oncogenic activity and has potential to be a novel biomarker in numerous cancers.

Several microRNAs have been shown that when deregulated contribute to malignant phenotypes. The miR-210-3p participates in important signaling pathways, regulating DNA damage repair, cell cycle, cellular death, stem cell differentiation, immune response, angiogenesis and even mitochondrial metabolism [125]. MiR-210-3p is tightly regulated by hypoxia conditions, commonly found in solid tumors [125]. Indeed, miRNA-210-3p levels in glioblastoma, pancreatic, breast, lung and head and neck cancers are higher than in normal tissue [126]. Shao et al. [126] investigated the transcript's role in cervical cancer tissue samples and observed that miRNA-210-3p expression is highly correlated with tumor differentiation, lymphatic metastasis and FIGO (International Federation of Gynecology and Obstetrics) tumor staging. Similarly, miR-154 expression level was studied in several biological mechanisms and was found to be linked to cardiac complications, diabetes, diabetic kidney disease, spermatogenesis and endometritis [127]. MiR-154 also acts as a tumor suppressor silencing different gene sets in several cancers, regulating cell cycle arrest, EMT, apoptosis, proliferation, migration, metastasis and even sensitivity to doxorubicin treatment on breast cancer patients. Decreased expression of miR-154 is proposed to be a key element to tumorigenesis in breast cancer, colorectal cancer, glioma, hepatocellular carcinoma, prostate cancer, non-small cell lung cancer and gastric cancer. Moreover, the oncogenic lncRNAs SNHG1 (colorectal cancer), SNHG5 (breast cancer), SNHG20 (non-small cell lung cancer), PCNAP1 (hepatocellular carcinoma) and the circular RNA Circ_101064 (glioma) have a sponging effect on miR-154, acting as competitive endogenous RNAs [127]. The expression profiles of circulating ncRNAs, such as miRNAs, can vary upon different treatment steps, disease stage and post-surgery status. The lack of an established endogenous miRNA control and various extraction and quantification methods represents another major challenge to normalize data for circulating miRNAs levels [128].

In addition to proliferative diseases, there are multiple examples of ncRNAs associated to degenerative, inflammatory, cardiovascular and syndromic diseases, among other (Table 5.3).

Table 5.3 Examples of ncRNAs involved in different human pathologies

Disease	Name	Class	References
Chr5q syndrome	miR-145 and miR-146a	miRNA	[129]
Alzheimer's disease	miR-29, miR-146 and miR-107	miRNA	[130–132]
	ncRNA antisense transcript for BACE1	lncRNA	[133]
Amyotrophic lateral sclerosis	miR143-3p, miR-206, miR-208b, miR-374b-5p, miR-499	miRNA	[134]
	NEAT1	lncRNA	[135, 136]
Arrhytmia and hypertension	miR-1	miRNA	[137]
Atheromatosis and atherosclerosis	miR-10a, miR-145, mR-143 and miR-126	miRNA	[138–140]
	Circular ncRNA linked to the CDKN2A locus	lncRNA	[141]
Beckwith–Wiedeman syndrome	lncRNAs H19 and KCNQ1OT1	lncRNA	[142]

(continued)

Table 5.3 (continued)

Disease	Name	Class	References
Bladder cancer	miR-205	miRNA	[143]
Breast cancer	U50	snoRNA	[144]
	miR-200c, miR-141, miR-148a	miRNA	[145–147]
	Uc.160+, Uc.283+A, Uc.346+	T-UCR	[113, 148]
	HOTAIR, TINCR, LINC00511, PPP1R26-AS1	lncRNA	[106, 149]
Cardiac hypertrophy	miR-21	miRNA	[150]
Colon cancer	miR-200c, miR-141, miR-129-2, miR-124a, miR-148a, miR-9, miR-137	miRNA	[145–147, 151–154]
	Uc.160+, Uc.283+A, Uc.346+	T-UCR	[113, 148]
Crohn's disease	miR-196	miRNA	[155]
Cystic fibrosis	miR-9, miR-93, miR-145-5p, miR-181b, miR-454, miR-509-3p	miRNA	[156–164]
	XIST, TLR8, HOTAIR, MALAT1, TLR8-AS1, BLACAT1, MEG9, BGas	lncRNA	[165–167]
Deafness	miR-96	miRNA	[168]
Down's syndrome	mir-155 and miR-802	miRNA	[169]
Duchenne muscular dystrophy	miR-1, miR-133, miR-206	miRNA	[170, 171]
	lnc-31, linc-MD1	lncRNA	[172–174]
Endometrial cancer	miR-129-2	miRNA	[153]
Facioscapulohumeral muscular dystrophy	miR-411	miRNA	[175]
	DBE-T	lncRNA	[176]
Familial dysautonomia	miR-203a-3p	miRNA	[177]
Gastric cancer	miR-124a, miR-129-2, miR-196b	miRNA	[178, 179]
Hailey–Hailey disease	miR-99a, miR-106, miR-125b, miR-181a	miRNA	[180]
Head and neck cancer	miR-137, miR-9	miRNA	[151, 181]
Hepatocellular carcinoma	miR-99a, miR-210	miRNA	[182, 183]
ICF syndrome	miR-34b, miR-34c, miR-99b, let-7e and miR-125a	miRNA	[184]
Idiopathic neurodevelopmental disease	T-UCRs uc.195, uc.392, uc.46 and uc.222	T-UCR	[185]
Lesch–Nyhan disease	miR-9, miR-181a, miR-187, miR-424	miRNA	[186]
Leukaemia	Uc.159, Uc.21, Uc.72	T-UCR	[187]
Li-Fraumeni syndrome	miR-605	miRNA	[188]
Lung cancer	miR-200c, miR-141	miRNA	[145, 146]
	Uc.160+, Uc.283+A, Uc.346+	T-UCR	[113, 148]
McCune–Albright syndrome	lncRNA NESP-AS	lncRNA	[142]
Melanoma	let-7a and b, miR-148, miR-155, miR-182, miR-200c, miR-211, miR-214, miR-221, miR-222	miRNA	[189]
MELAS syndrome	miR-9	miRNA	[190]
	LINC01405, SNHG12, RP11-403P17.4, CTC-260E6.6, RP11-357D18.1	lncRNA	[190]

Table 5.3 (continued)

Disease	Name	Class	References
Multiple osteochondromas	miR-21, miR-140, miR-145, miR-195, miR-214, miR-451, miR-483	miRNA	[191]
Myotonic dystrophy (type 1)	miR-1, miR-133a/b, miR-206	miRNA	[192, 193]
	MALAT1	lncRNA	[194]
Neuroectodermal brain tumors	miR-517c and miR-520 g	miRNA	[195]
Oesophageal adenocarcinoma	miR-106b-25	miRNA	[196]
Pancreatic cancer	miR-9, miR-124a	miRNA	[197, 198]
	AF339813, AFAP1-AS1, BC008363, ENST00000480739, GAS5, H19, HOTAIR, HOTTIP, MALAT-1, PVT1	lncRNA	[199]
Parkinson's disease	miR-7, miR-184 and let-7	miRNA	[200]
Prader–Willi and Angelman syndromes	snoRNA cluster at 15q11–q13 imprinted locus	snoRNA	[201]
Pseudohypoparathyroidism	lncRNA NESP-AS	lncRNA	[202]
Pulmonary arterial hypertension	miR-9, miR-124, miR-130, miR-206	miRNA	[203]
	MEG3, LnRPT	lncRNA	[204, 205]
Rett syndrome	miR-29b, miR-92, miR-122a, miR-130, miR-146a, miR-146b, miR-199a, miR-199b, miR-221, miR-296, miR-329, miR-342, miR-382, miR-409,	miRNA	[206, 207]
	AK081227, AK087060	lncRNA	[208]
Rheumatoid arthritis	miR-146a	miRNA	[209]
Sézary syndrome	miR-18a, miR-21, miR-31, miR-199a2, miR-214, miR-233, miR-342, miR-486	miRNA	[210, 211]
Silver–Russell syndrome	miR-675	miRNA	[142]
	lncRNA H19	lncRNA	[142]
Spinal motor neuron disease	miR-9	miRNA	[212]
Spinocerebellar ataxia type 1	miR-19, miR-101, miR-100	miRNA	[213]
Transient neonatal diabetes mellitus	lncRNA HYMAI	lncRNA	[214]
Ullrich congenital muscular dystrophy	miR-30c, miR-181a	miRNA	[215]
Uniparental disomy 14	snoRNA cluster at 14q32.2 imprinted locus	snoRNA	[142]
β-Thalassemia	miR-15a, miR-16-1, miR-26b, miR-96, miR-144, miR-155, miR-181a/c, miR-210, miR-320, miR-451, miR-486-3p, miR-503	miRNA	[216–222]
	DQ583499, XIST, lincRNA-TPM1, MRFS16P, lincRNA-RUNX2–2, HMI-LNCRNA, NR_001589, NR_120526, T315543	lncRNA	[223–225]

As an example not related to cancer, [226] analyzed lncRNAs of plasma exosomes isolated from coronary artery disease (CAD) patients and controls and found that the circulating exosomal lncRNA SOCS2-AS1 levels were negatively correlated with platelet and lipoprotein(a), both known to be associated with CAD. Therefore, high expression of SOCS2-AS1 was considered to be a protective factor against CAD and has potential to be a novel biomarker for coronary artery disease. Several studies have explored miRNAs as biomarker of coronary diseases. Karakas et al. measured eight miRNAs expression levels in more than a thousand CAD patients for four years and found circulating miR-132, miR-140-3p, and miR-21 to be independent predictors of cardiovascular death [227]. The three miRNAs have a better prognostic power in a subgroup of individuals with acute coronary syndrome (ACS) comorbidity. Schulte et al. found two miRNAs (platelet-related miR-197 and miR-223) as predictors of cardiovascular death in CAD patients, also with greater precision in previous ACS patients [228]. The circulating miR-26b-5p, miR-320a and miR-660-5p were found to be prognostic biomarkers for cardiac death or recurrent myocardial infarction in STEMI (ST-segment elevation myocardial infarction) patients [229]. A 10-year follow-up study focused on acute myocardial infarction (AMI) occurrence proposed that circulating miR-126, miR-197 and miR-223 could be prognostic biomarkers of AMI [230]. Surprisingly, Jansen et al. [231] found that increased miR-126 and miR-199a levels in circulating endothelial- and platelet-derived microvesicles (but not freely circulating in plasma) in CAD patients were associated with lower cardiovascular event rates [231]. The discordance of the miR-126's role could be explained by the fact that activated platelets-released miRs regulate adhesion molecules expression in endothelial cells [232]. Therefore, microvesicle encapsulation might change the miRs functions in intercellular signaling.

Single-nucleotide polymorphism (SNP) are common (>1%) genetic variants [233]. A simple alternative nucleotide may result in premature stop codon, nonsynonymous substitution in the translated protein, loss of stop codon, frameshift or even have no impact on the final protein sequence. SNPs may also alter gene expression if present in regions of promoter or binding of inhibitory complexes [234]. Besides generating phenotypic variability, SNPs may be responsible for certain disease susceptibilities [235]. More than 75% of rare single nucleotide variants (SNV) disclosed as pathogenic occur at a specific position in a given genome and most are located in noncoding promoter and enhancer regions [233]. Although the largest impact of SNVs has been studied in protein-coding sequences, there are several cases of SNP/SNVs occurring in noncoding regions. In fact, less than 10% of cancer-related SNPs map to protein-coding sequences [233]. Also, SNPs in lincRNAs can alter its biochemical characteristics, such as folding, binding partners, stability and regulatory networks [236]. Importantly, several SNPs in ncRNAs have been associated with clinicopathological features and proposed as prognostic biomarkers (Table 5.4, from [233]). The lincRNA *LINC00860*/*CASC8* (Cancer susceptibility 8) gene harbors SNV rs378854, related to adiposity in African-descent individuals; and rs10505477 and rs7837328, which correlate with higher lung, colorectal and gastric cancer risk [237–239]. The Rs217727 variant regulates the

Table 5.4 Trait-associated single-nucleotide variants in lncRNAs (from [233])

LncRNA	Trait-associated variants	Diseases	Position	References
CASC8	rs378854	Adiposity	Intron	[237]
	rs10505477	Colorectal, gastric, and lung cancers	Intron	[238, 239]
CASC19	rs138042437	Prostate cancer	Intron	[240]
CCAT1	rs6983267	Colorectal cancer, endometrial carcinoma	Enhancer	[241, 242]
CCAT2	rs6983267	Prostate, breast, colon, and colorectal cancers; myeloid malignancies	Exon	[243–246]
PCAT1	rs7463708	Prostate cancer	Enhancer	[247]
	rs10086908	Prostate cancer	Promoter	[247]
PRNCR1	rs1456315, rs7463708	Prostate cancer	Exon	[248]
	rs13252298, rs1456315	Colorectal cancer	Exon	[249]
	rs183373024	Prostate cancer	Exon	[240]
PVT1	rs13281615	Breast cancer	Promoter	[250]
	rs2720709, rs2648875	End-stage renal disease (ESRD)	Intron, intron	[251]
	rs378854	Prostate cancer	Promoter	[252]
	rs13255292, rs4733601	Diffuse large B cell lymphoma	Intron, downstream	[253]
CASC16	rs3803662	Breast cancer, lung cancer	Exon	[254]
CASC15	rs6939340	Neuroblastoma	Intron	[255]
GAS5	rs145204276	Hepatocellular carcinoma (HCC), colorectal, and gastric cancers	Promoter	[256, 257]
H19	rs217727	Coronary artery disease, type 2 diabetes	Exon	[258]
	rs2067051	Pneumoconiosis, coronary artery disease	Exon	[258, 259]
	rs2107425	Ovarian and breast cancers, hypertrophic cardiomyopathy	Intron	[260]
	rs2839698	HCC, bladder, colorectal, and gastric cancer	Exon	[260–262]
HULC	rs7763881, rs1041279	HCC	Intron	[263]
LINC00673	rs11655237	Pancreatic cancer	Exon	[264]
LINC00951	rs11752942	Esophageal squamous cell carcinoma (ESCC)	Exon	[265]
LOC105378318	rs1875147	Leprosy	Intron	[266]

(continued)

Table 5.4 (continued)

LncRNA	Trait-associated variants	Diseases	Position	References
MALAT1	rs619586	Pulmonary arterial hypertension (PAH), coronary atherosclerotic and congenital heart disease (CAD/CHD), breast cancer	Exon	[267, 268]
	rs1194338	Colorectal cancer	Promoter	[269]
	rs4102217	HCC	Promoter	[270]
MEG3	rs941576, rs34552516	Type 1 diabetes (T1D)	Intron	[271, 272]
MIAT	rs2331291	Myocardial infarction	Intron	[273]
	rs1894720	Paranoid schizophrenia	Exon	[274]
PCGEM1	rs6434568, rs16834898	Prostate cancer	Intron	[275]
PCAT19	rs11672691	Prostate cancer	Promoter	[276]
PTCSC2	rs965513	Papillary thyroid carcinoma (PTC)	Intron	[277]
PTCSC3	rs944289	PTC, large-vessel ischemic stroke	Promoter	[278, 279]
TDRG1	rs8506	ESCC, gastric cancer	Exon	[280]
TINCR	rs2288947, rs8105637	Colorectal cancer, gastric cancer	Exon, intron	[281]

expression of its host sequence, lincRNA *H19* (H19 imprinted maternally expressed transcript), promoting apoptosis in hepatocellular carcinoma [282]. Also in hepatocellular carcinoma, the SNPs rs2839698, rs2735971 and rs3024270 in *H19* could be markers for a higher cancer risk, while rs2839698 also associates with poor prognosis [262]. In contrast, *H19* SNP rs2067051 was associated with a decreased risk for pneumoconiosis in coal workers in a Chinese population study [259]. The SNVs rs2720709 and rs2648875 found in the lncRNA *PVT1* oncogene are related with development of end-stage renal disease (ESRD) in type 2 diabetes patients [251]. Also in lincRNA *PVT1*, a SNV alone or in epistatic interaction with SNPs in other risk genes might help predict an optimal response to glatiramer acetate treatment for multiple sclerosis [283]. A genome-wide association study (GWAS) conducted by Cerhan et al. [253] observed that *PVT1* rs13255292 and rs4733601 are risk SNPs with predictive value for susceptibility of diffuse large B cell lymphoma. Ghesquières et al. [284] found that rs2608053 in *PVT1* is associated with survival outcome in cases of Hodgkin lymphoma. In prostate cancer, rs7463708 T genotype up-regulates its host lincRNA, *PCAT1* (Prostate cancer associated transcript 1), by increasing binding of the androgen receptor (AR)-interacting transcription factor ONECUT2 (one cut homeobox 2) to a *PCAT1*-associated enhancer. Therefore, rs7463708 could be a biomarker for prostate cancer susceptibility [247]. In hepatocellular carcinoma, lincRNA *MALAT1* (Metastasis-associated lung adenocarcinoma transcript 1) SNV rs4102217 is associated with cancer risk [270]. However, *MALAT1* rs3200401 T allele was associated with longer survival in lung adenocarcinoma cases [285].

Besides its prognostic features, SNP/SNVs may also be used as markers for assessing a patient's response to treatment. Upon platinum-based chemotherapy for lung cancer, Gong et al. [286] observed association between lincRNA *MEG3* (Maternally expressed 3) SNV rs116907618 and severe gastrointestinal toxicity; and lincRNA *CDKN2B-AS1*/*ANRIL* (CDKN2B antisense RNA 1) rs1333049 and overall toxicity (mainly severe hematologic and gastrointestinal). For the same treatment, Hu et al. [239] observed that *CASC8*/*LINC00860* SNV rs10505477 was associated with severe hematologic toxicity in non-small-cell lung cancer and gastrointestinal toxicity in small-cell lung cancer. It is worth noting, that *CASC8* rs10505477 was found to be associated with several cancers, including invasive ovarian, colorectal and gastric cancer [239]. In gastric cancer, *CASC8* rs10505477 GG genotype might be associated with longer patient survival and tumor size, tumor–node–metastasis stage and lymph node metastasis [238].

In acute myeloid leukemia, *GAS5* (growth arrest specific 5) promoter SNP rs55829688 C allele up-regulates the lncRNA host by interacting with the transcription factor TP63, conferring a poor prognosis [287]. Guo et al. [288] found that *GAS5* SNP rs2067079 is associated with risk of neutropenia and severe myelosuppression upon chemoradiotherapy for nasopharyngeal carcinoma. However, the same study discovered that *GAS5* rs6790 might have a protective effect, reducing the toxic reaction rate to treatment.

Most of disease susceptibility risk loci found through GWAS map in noncoding regions with regulatory features [289]. Identification of SNPs in noncoding sequences represent a special challenge compared to variants in protein-coding regions and demand specific bioinformatic pipelines. The initial screening for genetic variants can be performed using data and annotations generated from public consortia such as the 1000 Genomes Project [290] and the NHGRI-EBI GWAS Catalog [291]. The latter is a systematic effort to summarize human SNP-trait associations found in several published and unpublished GWAS data. SNPs can be associated with phenotypic traits by expression quantitative trait loci (eQTL; see Chap. 4) analysis, for example. Another method is to identify putative SNPs that occur in regions of binding of regulatory molecules. In the latter example, it is useful to analyze transcription factor binding sites annotated in different datasets. The GWAS4D integrate GWAS data and tissue or cell type epigenomic profiles to identify context-specific regulatory genetic variants [289]. The dbSNP aggregates data of SNPs, microsatellites and small-scale deletions and insertions [292]. LincSNP (V. 3.0; http://bio-bigdata.hrbmu.edu.cn/lincsnp/) annotates disease or phenotype-associated genetic variants, including SNPs, in lncRNAs and circRNAs or their regulatory sequences, including transcription factor binding sites. LncRNASNP2 (http://bioinfo.life.hust.edu.cn/lncRNASNP#!/) integrates SNVs and mutation data with structural data of human and mouse lncRNAs and effects on miRNA binding features [293]. LncVar (http://159.226.118.31/LncVar/) focus on genetic variants and their impact on the biological features of its host lncRNAs in several species [294].

5.5 Challenges and Open Questions in the Noncoding RNA Field

Technological developments and application of NGS methods have paved the way for the investigation of the transcriptional landscape in human tissues and cell at an unprecedented resolution, leading to the discovery of several novel types of RNAs whose primary function is not to encode proteins. However, several challenges and opportunities still lay ahead for the full appreciation of the nature and biological significance of these molecules. Noncoding RNAs are typically expressed at lower levels and show greater tissue specificity than protein-coding genes, which poses a difficulty for the generation of a complete catalogue of the ncRNAs expressed in different cell types based on transcriptome sequencing of bulk or even microdissected tissue samples. Single cell RNAseq (scRNA-seq) methods can pottentially overcome this limitation as illustrated by recent studies that used scRNA-seq analysis to identify lncRNAs signatures associated to functionally distinct T cell subtypes isolated from human tumor [295] or cells from human neocortex at different stages of development [296].

Also, conventional sequencing by synthesis-based RNA-seq methods [297] require the conversion by reverse transcription (RT) of polyadenylated and/or nonpolyadenylated RNA populations into cDNA prior to NGS sequencing. The RT step precludes the direct detection of chemical modifications in nucleotide bases resulting from post-transcriptional RNA editing, the most frequent being the methylation of adenosines at the N^6-position (m6A) at specific sequence motifs by a nuclear RNA methyltransferase complex [298], and hydrolytic deamination of adenosines to inosines (A-I) catalyzed by members of the adenosine deaminase (ADAR) family [299]. A-I and m6A are frequent modifications found in protein-coding mRNAs and ncRNAs alike. The m6A modification at specific sites alters structural properties of the molecule. In protein-coding mRNAs it can affect splicing, stability and nuclear export thereby modulating gene expression. In ncRNAs, the m6A modification has been shown to affect lincRNA XIST stability and the ability of lincRNA MALAT-1 to interact with nuclear hnRNP particles, as well as processing of pre-miRNA precursors affecting miRNA biogenesis [300]. A-I edition may alter protein sequence by change in codon sequences or by creating/destroying mRNA splicing sites thereby affecting constitutive and alternative splicing. It has been detected in both small and long ncRNAs and structural changes induced by A-I possibly modify their biogenesis, structure and function [299]. Direct RNA sequencing based on the Oxford nanopore platform performs RNA sequencing without the necessity of the reverse transcription step thus allowing the detection of the modifications present in the primary RNA sequence [301]. Using this technology, it is possible to map the precise composition and relative abundance of RNA modifications in noncoding RNAs in different cell types and tissues and its dynamics in developmental states and pathological conditions, and obtain relevant information to establish the biological significance of these modifications.

Noncoding RNAs modulate gene expression as component parts of ribonucleoprotein complexes that associate with DNA and nucleosomes to chromatin activation states. Thus, revealing the complex network of RNA–chromatin associations is essential for understanding the tridimensional organization of the genome and its implication on global and local gene expression patterns. To address this it will be required to integrate in the same cell type high-resolution methodologies as tissue/cell specific transcriptional maps and chromatin accessibility maps. The latter maps also need to be generated by different methods as DNase I hypersensitive site (DHS) assays [302] and/or ATAC-seq [303] and chromosome conformation capture (3C)-related technologies [304] to probe long-range physical interactions between chromosome regions that can be separated by hundreds of kilobases in the linear sequence. Further integration with more recently developed direct approaches to globally map RNA-to-DNA contacts across the genome will increase the resolution and improve the specificity to identify specific RNA–chromatin interactions that contribute to gene expression regulation. These are based on proximity ligation followed by sequencing such as Chromatin-Associated RNA sequencing (ChAR-seq) [305] or global RNA interactions with DNA by deep sequencing (GRID-seq) [306].

As tons of data are being generated through large-scale sequencing projects, evidence of functionality of ncRNAs accumulate exponentially, but the pace at which this information is translated into knowledge about the underlying molecular mechanisms has been much slower. A key limitation to facilitate the study of RNA functions not associated with mRNA translation is that the RNA primary sequence does not readily inform about structure-function relationships. The generally low level of evolutionary conservation of ncRNAs limits the power of comparative genomic analyses to identify sequence motifs under purifying selection that would point to potentially functional domains in ncRNA sequences [12]. On the other hand, it has been noted through interspecies comparative sequence analysis that many ncRNAs, despite the primary sequence conservation, maintain positional syntenic conservation (i.e., maintain the same relative genomic position to neighboring genes or regulatory elements in evolutionarily related species). Also, the analysis of nucleotide covariation patterns through comparative multi-sequence alignments predict conserved secondary structures and domains that further corroborate the existence of positive selection acting on specific ncRNAs. To achieve a better knowledge of the structural spatial arrangement of RNA secondary domains and the dynamics upon interaction with RNA binding proteins it will be absolutely required to integrate data from biochemical methods that map intramolecular RNA-RNA interactions and secondary structure formation with biophysical methods such as NMR that inform about the dynamics of ribonucleoprotein complexes. Once the rules defining the spatial structure of RNA domains are known, they still need to be assigned to molecular or cellular phenotype to demonstrate functional relevance in loss/gain of function experiments. RNA interference using transfected siRNAs or recombinant shRNA expression have been extensively used to knock down specific ncRNAs and study the resulting phenotype. Unfortunately, siRNA approaches have limited efficacy against nuclear transcripts, being more useful to analyze cytoplasmic RNAs and post-transcriptional RNA-dependent process. Antisense oligonucleotides (ASO) induce RNase H activity and were shown to be more adequate to study nuclear ncRNAs, including *in vivo* experiments [307]. In a more robust and scalable approach, Liu et al. [308] generated a CRISPR interference library, based on a nuclease-dead dCas9 fused to a repressive KRAB domain, expressing sgRNAs that target promoters of more than 16,000 human lncRNAs. This approach induces gene silencing through deposition of repressive HeK27me3 marks on promoters and was used to screen for noncoding RNAs essential for cell growth [308]. The study revealed approximately 500 lncRNAs that are essential for cell survival and growth, most of which are specific for only one cell type, highlighting the selectivity of lncRNAs expression.

Finally, many lncRNAs harbor small open reading frames (smORFs) that potentially produce peptides with less than 100 amino acids. A fraction of these RNAs is associated to ribosomes in polysome profiling RNA-seq experiments, and in these cases it is challenging to rule out without further experimentation that the biologically relevant functions of these molecules is to encode polypeptides [37]. A way proposed to overcome this conundrum is to analyze the ratio between synonymous vs. non-synonymous substitutions in the lncRNA ORF as a proxy codon sequence preservation, but this approach is limited by the low level of evolutionary conservation that precludes the selection of orthologous sequences for sequence analysis [309]. In fact, smORF sequences show higher conservation than introns in general, but lower than the observed in mRNAs [56, 310]. The smORFs may potentially drive the synthesis of small peptides, which bring a new layer of complexity in the analysis of ncRNAs [56, 311]. The integration of *in silico* smORF prediction in ncRNAs retrieved from data generated by combined RNA-seq and mass spectrometry of ribosome profiling can be a solid strategy to train computational machine learning-based prediction methods to more effectively discriminate bonafide ncRNAs from those with smORFs that may undergo efficient translation [312, 313].

References

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. https://doi.org/10.1038/35057062.
2. Litwack G (2018) Nucleic acids and molecular genetics. In: Human biochemistry. Elsevier, pp. 257–317.
3. Feingold EA, Good PJ, Guyer MS, et al. The ENCODE (ENCyclopedia of DNA Elements) Project. Science. 2004;306:636–40.
4. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74. https://doi.org/10.1038/nature11247.
5. Moore JE, Purcaro MJ, Pratt HE, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710. https://doi.org/10.1038/s41586-020-2493-4.
6. Doolittle WF. We simply cannot go on being so vague about "function". Genome Biol. 2018;19:223. https://doi.org/10.1186/s13059-018-1600-4.

7. Westheimer FH. Biochemistry: Polyribonucleic acids as enzymes. Nature. 1986;319:534–6. https://doi.org/10.1038/319534a0.
8. Cech TR. Self-splicing RNA: implications for evolution. Int Rev Cytol. 1985;93:3–22. https://doi.org/10.1016/S0074-7696(08)61370-4.
9. Cech TR, Steitz JA. The noncoding RNA revolution - trashing old rules to forge new ones. Cell. 2014;157:77–94.
10. Turner, Douglas H., Mathews DH (ed) (2016) RNA structure determination.
11. Low JT, Weeks KM. SHAPE-directed RNA secondary structure prediction. Methods. 2010;52:150–8.
12. Rinn JL, Chang HY, Chang HY. Long noncoding RNAs: molecular modalities to organismal functions. Annu Rev Biochem. 2020;89:283–308.
13. Schaefer M, Kapoor U, Jantsch MF. Understanding RNA modifications: the promises and technological bottlenecks of the "epitranscriptome". Open Biol. 2017;7:170077.
14. Roundtree IA, Evans ME, Pan T, He C. Dynamic RNA modifications in gene expression regulation. Cell. 2017;169:1187–200.
15. Kim H-J, Jeong S-H, Heo J-H, et al. mRNA capping enzyme activity is coupled to an early transcription elongation. Mol Cell Biol. 2004;24:6184–93. https://doi.org/10.1128/mcb.24.14.6184-6193.2004.
16. St.Laurent G, Wahlestedt C, Kapranov P. The landscape of long noncoding RNA classification. Trends Genet. 2015;31:239–51.
17. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22:1775–89. https://doi.org/10.1101/gr.132159.111.
18. Eddy SR. Noncoding RNA genes. Curr Opin Genet Dev. 1999;9:695–9.
19. Telonis AG, Loher P, Kirino Y, Rigoutsos I. Nuclear and mitochondrial tRNA-lookalikes in the human genome. Front Genet. 2014;5:344. https://doi.org/10.3389/fgene.2014.00344.
20. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016;44:D184–9. https://doi.org/10.1093/nar/gkv1309.
21. Malinovskaya EM, Ershova ES, Golimbet VE, et al. Copy number of human ribosomal genes with aging: unchanged mean, but narrowed range and decreased variance in elderly group. Front Genet. 2018;9:306. https://doi.org/10.3389/fgene.2018.00306.
22. Agrawal S, Ganley ARD. The conservation landscape of the human ribosomal RNA gene repeats. PLoS One. 2018;13:e0207531. https://doi.org/10.1371/journal.pone.0207531.
23. Bratkovič T, Bozič J, Rogelj B. Functional diversity of small nucleolar RNAs. Nucleic Acids Res. 2020;48:1627–51. https://doi.org/10.1093/nar/gkz1140.
24. Hombach S, Kretz M. Non-coding RNAs: classification, biology and functioning. Adv Exp Med Biol. 2016;937:3–17. https://doi.org/10.1007/978-3-319-42059-2_1.
25. Kosmyna B, Gupta V, Query C (2020) Transcriptional analysis supports the expression of human snRNA variants and reveals U2 snRNA homeostasis by an abundant U2 variant. bioRxiv 2020.01.24.917260.
26. Webb CJ, Zakian VA. Telomerase RNA is more than a DNA template. RNA Biol. 2016;13:683–9. https://doi.org/10.1080/15476286.2016.1191725.
27. Egan ED, Collins K. An enhanced H/ACA RNP assembly mechanism for human telomerase RNA. Mol Cell Biol. 2012;32:2428–39. https://doi.org/10.1128/mcb.00286-12.
28. Rosenblad MA, Larsen N, Samuelsson T, Zwieb C. Kinship in the SRP RNA family. RNA Biol. 2009;6:508–16. https://doi.org/10.4161/rna.6.5.9753.
29. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011;12:861–74.
30. Taft RJ, Kaplan CD, Simons C, Mattick JS. Evolution, biogenesis and function of promoter-associated RNAs. Cell Cycle. 2009;8:2332–8.
31. Kapranov P, Cheng J, Dike S, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007;316:1484–8. https://doi.org/10.1126/science.1138341.

32. Seila AC, Calabrese JM, Levine SS, et al. Divergent transcription from active promoters. Science. 2008;322:1849–51. https://doi.org/10.1126/science.1162253.
33. Yu D, Ma X, Zuo Z, et al. Classification of transcription boundary-associated RNAs (TBARs) in animals and plants. Front Genet. 2018;9:168.
34. Xu Z, Wei W, Gagneur J, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009;457:1033–7. https://doi.org/10.1038/nature07728.
35. Preker P, Almvig K, Christensen MS, et al. PROMoter uPstream transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res. 2011;39:7179–93. https://doi.org/10.1093/nar/gkr370.
36. Ntini E, Järvelin AI, Bornholdt J, et al. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol. 2013;20:923–8. https://doi.org/10.1038/nsmb.2640.
37. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009;10:155–9.
38. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458:223–7. https://doi.org/10.1038/nature07672.
39. Ørom UA, Derrien T, Guigo R, Shiekhattar R. Long noncoding RNAs as enhancers of gene expression. Cold Spring Harb Symp Quant Biol. 2010;75:325–31. https://doi.org/10.1101/sqb.2010.75.058.
40. Huarte M. The emerging role of lncRNAs in cancer. Nat Med. 2015;21:1253–61. https://doi.org/10.1038/nm.3981.
41. Fabris L, Calin GA. Understanding the genomic ultraconservations: T-UCRs and cancer. Int Rev Cell Mol Biol. 2017;333:159–72.
42. Jarroux J, Morillon A, Pinskaya M. History, discovery, and classification of lncRNAs. Adv Exp Med Biol. 2017;1008:1–46.
43. Nagai K, Oubridge C, Kuglstatter A, et al. Structure, function and evolution of the signal recognition particle. EMBO J. 2003;22:3479–85.
44. Valadkhan S, Gunawardane LS. Role of small nuclear RNAs in eukaryotic gene expression. Essays Biochem. 2013;54:79–90. https://doi.org/10.1042/BSE0540079.
45. Filipowicz W, Pogači V. Biogenesis of small nucleolar ribonucleoproteins. Curr Opin Cell Biol. 2002;14:319–27.
46. Richard P, Darzacq X, Bertrand E, et al. A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs. EMBO J. 2003;22:4283–93. https://doi.org/10.1093/emboj/cdg394.
47. Egloff S, Studniarek C, Kiss T. 7SK small nuclear RNA, a multifunctional transcriptional regulatory RNA with gene-specific features. Transcription. 2018;9:95–101.
48. Fire A, Xu S, Montgomery MK, et al. Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. Nature. 1998;391:806–11. https://doi.org/10.1038/35888.
49. Moyano M, Stefani G. PiRNA involvement in genome stability and human cancer. J Hematol Oncol. 2015;8:1–10. https://doi.org/10.1186/s13045-015-0133-5.
50. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215–33. https://doi.org/10.1016/j.cell.2009.01.002.
51. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 2007;8:93–103.
52. Sætrom P, Heale BSE, Snøve O, et al. Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res. 2007;35:2333–42. https://doi.org/10.1093/nar/gkm133.
53. Henriques T, Gilchrist DA, Nechaev S, et al. Stable pausing by rna polymerase II provides an opportunity to target and integrate regulatory signals. Mol Cell. 2013;52:517–28. https://doi.org/10.1016/j.molcel.2013.10.001.

54. Preker P, Nielsen J, Kammler S, et al. RNA exosome depletion reveals transcription upstream of active human promoters. Science. 2008;322:1851–4. https://doi.org/10.1126/science.1164096.
55. Taft RJ, Hawkins PG, Mattick JS, Morris KV. The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization. Epigenet Chromatin. 2011;4:13. https://doi.org/10.1186/1756-8935-4-13.
56. Ransohoff JD, Wei Y, Khavari PA, et al. The functions and unique features of long intergenic non-coding RNA HHS Public Access. Nat Rev Mol Cell Biol. 2018;19:143–57. https://doi.org/10.1038/nrm.2017.104.
57. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet. 2014;15:7–21. https://doi.org/10.1038/nrg3606.
58. Kopp F, Mendell JT. Functional classification and experimental dissection of long noncoding RNAs. Cell. 2018;172:393–407. https://doi.org/10.1016/J.CELL.2018.01.011.
59. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66. https://doi.org/10.1146/annurev-biochem-051410-092902.
60. Rinn JL, Kertesz M, Wang JK, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007;129:1311–23. https://doi.org/10.1016/j.cell.2007.05.022.
61. Zhao J, Ohsumi TK, Kung JT, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell. 2010;40:939–53. https://doi.org/10.1016/j.molcel.2010.12.011.
62. Wang KC, Yang YW, Liu B, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011;472:120–6. https://doi.org/10.1038/nature09819.
63. Yang YW, Flynn RA, Chen Y, et al. Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. Elife. 2014;2014:2046. https://doi.org/10.7554/eLife.02046.001.
64. Cheng Y, Jutooru I, Chadalapaka G, et al. The long non-coding RNA HOTTIP enhances pancreatic cancer cell proliferation, survival and migration. Oncotarget. 2015;6:10840–52. https://doi.org/10.18632/oncotarget.3450
65. Ruthenburg AJ, Allis CD, Wysocka J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Mol Cell. 2007;25:15–30.
66. Zhao J, Sun BK, Erwin JA, et al. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science. 2008;322:750–6. https://doi.org/10.1126/science.1163045.
67. Ogawa Y, Sun BK, Lee JT. Intersection of the RNA interference and X-inactivation pathways. Science. 2008;320:1336–41. https://doi.org/10.1126/science.1157676.
68. Amaral PP, Mattick JS. Noncoding RNA in development. Mamm Genome. 2008;19:454–92.
69. Latos PA, Pauler FM, Koerner MV, et al. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science. 2012;338:1469–72. https://doi.org/10.1126/science.1228110.
70. Andergassen D, Muckenhuber M, Bammer PC, et al. The Airn lncRNA does not require any DNA elements within its locus to silence distant imprinted genes. PLoS Genet. 2019;15:e1008268. https://doi.org/10.1371/journal.pgen.1008268.
71. Yotova IY, Vlatkovic IM, Pauler FM, et al. Identification of the human homolog of the imprinted mouse Air non-coding RNA. Genomics. 2008;92:464–73. https://doi.org/10.1016/j.ygeno.2008.08.004.
72. Galganski L, Urbanek MO, Krzyzosiak WJ. Nuclear speckles: molecular organization, biological function and role in disease. Nucleic Acids Res. 2017;45:10350–68.
73. Sunwoo H, Dinger ME, Wilusz JE, et al. Men ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res. 2009;19:347–59. https://doi.org/10.1101/gr.087775.108.
74. Prasanth KV, Prasanth SG, Xuan Z, et al. Regulating gene expression through RNA nuclear retention. Cell. 2005;123:249–63. https://doi.org/10.1016/j.cell.2005.08.033.

75. Wight M, Werner A. The functions of natural antisense transcripts. Essays Biochem. 2013;54:91–101. https://doi.org/10.1042/BSE0540091.
76. Katayama S, Tomaru Y, Kasukawa T, et al. Molecular biology: Antisense transcription in the mammalian transcriptome. Science. 2005;309:1564–6. https://doi.org/10.1126/science.1112009.
77. Lam MTY, Li W, Rosenfeld MG, Glass CK. Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci. 2014;39:170–82.
78. Li W, Notani D, Ma Q, et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature. 2013b;498:516–20. https://doi.org/10.1038/nature12210.
79. Qu S, Yang X, Li X, et al. Circular RNA: a new star of noncoding RNAs. Cancer Lett. 2015;365:141–8.
80. Wang Y, Wang Z. Efficient backsplicing produces translatable circular mRNAs. RNA. 2015;21:172–9. https://doi.org/10.1261/rna.048272.114.
81. Sen R, Ghosal S, Das S, et al. Competing endogenous RNA: the key to posttranscriptional regulation. Sci World J. 2014;2014:896206.
82. Kartha RV, Subramanian S. Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. Front Genet. 2014;5:8.
83. Reis EM, Verjovski-Almeida S. Perspectives of long non-coding RNAs in cancer diagnostics. Front Genet. 2012;3:32. https://doi.org/10.3389/fgene.2012.00032.
84. Hayashita Y, Osada H, Tatematsu Y, et al. A polycistronic MicroRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. Cancer Res. 2005;65:9628–32. https://doi.org/10.1158/0008-5472.CAN-05-2352.
85. He L, Thomson JM, Hemann MT, et al. A microRNA polycistron as a potential human oncogene. Nature. 2005;435:828–33. https://doi.org/10.1038/nature03552.
86. Johnson SM, Grosshans H, Shingara J, et al. RAS is regulated by the let-7 microRNA family. Cell. 2005;120:635–47. https://doi.org/10.1016/j.cell.2005.01.014.
87. Takamizawa J, Konishi H, Yanagisawa K, et al. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. Cancer Res. 2004;64:3753–6. https://doi.org/10.1158/0008-5472.CAN-04-0637.
88. Eilebrecht S, Brysbaert G, Wegert T, et al. 7SK small nuclear RNA directly affects HMGA1 function in transcription regulation. Nucleic Acids Res. 2011;39:2057–72. https://doi.org/10.1093/nar/gkq1153.
89. Lanz RB, McKenna NJ, Onate SA, et al. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell. 1999;97:17–27. https://doi.org/10.1016/S0092-8674(00)80711-4.
90. Xiong T, Li J, Chen F, Zhang F. PCAT-1: a novel oncogenic long non-coding RNA in human cancers. Int J Biol Sci. 2019;15:847–56.
91. Poliseno L, Salmena L, Zhang J, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465:1033–8. https://doi.org/10.1038/nature09144.
92. Chu Y, Yue X, Younger ST, et al. Involvement of argonaute proteins in gene silencing and activation by RNAs complementary to a non-coding transcript at the progesterone receptor promoter. Nucleic Acids Res. 2010;38:7736–48. https://doi.org/10.1093/nar/gkq648.
93. Tsang WP, Ng EKO, Ng SSM, et al. Oncofetal H19-derived miR-675 regulates tumor suppressor RB in human colorectal cancer. Carcinogenesis. 2010;31:350–8. https://doi.org/10.1093/carcin/bgp181.
94. Li L, Feng T, Lian Y, et al. Role of human noncoding RNAs in the control of tumorigenesis. Proc Natl Acad Sci U S A. 2009a;106:12956–61. https://doi.org/10.1073/pnas.0906005106.
95. Kotake Y, Nakagawa T, Kitagawa K, et al. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15 INK4B tumor suppressor gene. Oncogene. 2011;30:1956–62. https://doi.org/10.1038/onc.2010.568.

96. Yap KL, Li S, Muñoz-Cabello AM, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell. 2010;38:662–74. https://doi.org/10.1016/j.molcel.2010.03.021.
97. Mourtada-Maarabouni M, Hasan AM, Farzaneh F, Williams GT. Inhibition of human T-cell proliferation by mammalian target of rapamycin (mTOR) antagonists requires noncoding RNA growth-arrest-specific transcript 5 (GAS5). Mol Pharmacol. 2010;78:19–28. https://doi.org/10.1124/mol.110.064055.
98. Huarte M, Guttman M, Feldser D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010;142:409–19. https://doi.org/10.1016/j.cell.2010.06.040.
99. Yochum GS, Cleland R, McWeeney S, Goodman RH. An antisense transcript induced by Wnt/β-catenin signaling decreases E2F4. J Biol Chem. 2007;282:871–8. https://doi.org/10.1074/jbc.M609391200.
100. Tazawa H, Tsuchiya N, Izumiya M, Nakagama H. Tumor-suppressive miR-34a induces senescence-like growth arrest through modulation of the E2F pathway in human colon cancer cells. Proc Natl Acad Sci U S A. 2007;104:15472–7. https://doi.org/10.1073/pnas.0707351104.
101. Andersson S, Wallin KL, Hellström AC, et al. Frequent gain of the human telomerase gene TERC at 3q26 in cervical adenocarcinomas. Br J Cancer. 2006;95:331–8. https://doi.org/10.1038/sj.bjc.6603253.
102. Sheedy P, Medarova Z. The fundamental role of miR-10b in metastatic cancer. Am J Cancer Res. 2018;8:1674–88.
103. Gregory PA, Bert AG, Paterson EL, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nat Cell Biol. 2008;10:593–601. https://doi.org/10.1038/ncb1722.
104. Nguyen HT, Li C, Lin Z, et al. The microRNA expression associated with morphogenesis of breast cancer cells in three-dimensional organotypic culture. Oncol Rep. 2012;28:117–26. https://doi.org/10.3892/or.2012.1764.
105. Liu X-H, Sun M, Nie F-Q, et al. Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. Mol Cancer. 2014;13:92. https://doi.org/10.1186/1476-4598-13-92.
106. Gupta RA, Shah N, Wang KC, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464:1071–6. https://doi.org/10.1038/nature08975.
107. Wang J, Liu X, Wu H, et al. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. Nucleic Acids Res. 2010;38:5366–83. https://doi.org/10.1093/nar/gkq285.
108. Würdinger T, Tannous BA, Saydam O, et al. miR-296 regulates growth factor receptor overexpression in angiogenic endothelial cells. Cancer Cell. 2008;14:382–93. https://doi.org/10.1016/j.ccr.2008.10.005.
109. Lu Z, Xiao Z, Liu F, et al. Long non-coding RNA HULC promotes tumor angiogenesis in liver cancer by up-regulating sphingosine kinase 1 (SPHK1). Oncotarget. 2016;7:241–54. https://doi.org/10.18632/ONCOTARGET.6280.
110. Chan JA, Krichevsky AM, Kosik KS. MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. Cancer Res. 2005;65:6029–33. https://doi.org/10.1158/0008-5472.CAN-05-0137.
111. Fu X, Ravindranath L, Tran N, et al. Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1. DNA Cell Biol. 2006;25:135–41. https://doi.org/10.1089/dna.2006.25.135.
112. Wing PT, Wong TWL, Cheung AHH, et al. Induction of drug resistance and transformation in human cancer cells by the noncoding RNA CUDR. RNA. 2007;13:890–8. https://doi.org/10.1261/rna.359007.

113. Calin GA, Chang-Gong L, Ferracin M, et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell. 2007;12:215–29. https://doi.org/10.1016/j.ccr.2007.07.027.
114. Khaitan D, Dinger ME, Mazar J, et al. The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. Cancer Res. 2011;71:3852–62. https://doi.org/10.1158/0008-5472.CAN-10-4460.
115. Puvvula PK, Desetty RD, Pineau P, et al. Long noncoding RNA PANDA and scaffold-attachment-factor SAFA control senescence entry and exit. Nat Commun. 2014;5:5323. https://doi.org/10.1038/ncomms6323.
116. Vincent K, Pichler M, Lee GW, Ling H. MicroRNAs, genomic instability and cancer. Int J Mol Sci. 2014;15:14475–91.
117. Lee S, Kopp F, Chang TC, et al. Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. Cell. 2016;164:69–80. https://doi.org/10.1016/j.cell.2015.12.017.
118. Lemos AEG, Da Rocha MA, Ferreira LB, Gimba ERP. The long non-coding RNA PCA3: an update of its functions and clinical applications as a biomarker in prostate cancer. Oncotarget. 2019;10:6589–603.
119. Ploussard G, Haese A, Van Poppel H, et al. The prostate cancer gene 3 (PCA3) urine test in men with previous negative biopsies: does free-to-total prostate-specific antigen ratio influence the performance of the PCA3 score in predicting positive biopsies? BJU Int. 2010;106:1143–7. https://doi.org/10.1111/j.1464-410X.2010.09286.x.
120. Zhan Y, Du L, Wang L, et al. Expression signatures of exosomal long non-coding RNAs in urine serve as novel non-invasive biomarkers for diagnosis and recurrence prediction of bladder cancer. Mol Cancer. 2018;17:142. https://doi.org/10.1186/s12943-018-0893-y.
121. Cui WC, Wu YF, Qu HM. Up-regulation of long non-coding RNA PCAT-1 correlates with tumor progression and poor prognosis in gastric cancer. Eur Rev Med Pharmacol Sci. 2017;21:3021–7.
122. Zhao X, Fan Y, Lu C, et al. PCAT1 is a poor prognostic factor in endometrial carcinoma and associated with cancer cell proliferation, migration and invasion. Bosn J Basic Med Sci. 2019;19:274–81. https://doi.org/10.17305/bjbms.2019.4096.
123. Song H, Liu Y, Jin X, et al. Long non-coding RNA LINC01535 promotes cervical cancer progression via targeting the miR-214/EZH2 feedback loop. J Cell Mol Med. 2019;23:6098–111. https://doi.org/10.1111/jcmm.14476.
124. Yang Y, Liu Y, Li G, et al. MicroRNA-214 suppresses the growth of cervical cancer cells by targeting EZH2. Oncol Lett. 2018b;16:5679–86. https://doi.org/10.3892/ol.2018.9363.
125. Pasculli B, Barbano R, Rendina M, et al. Hsa-miR-210-3p expression in breast cancer and its putative association with worse outcome in patients treated with Docetaxel. Sci Rep. 2019;9:1–9. https://doi.org/10.1038/s41598-019-51581-3.
126. Shao MX, Qu AZ, Wang YQ, Zhong YY. Expression level of miRNA-210-3p in cervical cancer and its prognostic potential. Eur Rev Med Pharmacol Sci. 2020;24:6583–8. https://doi.org/10.26355/eurrev_202006_21643.
127. Nazarizadeh A, Mohammadi F, Alian F, et al. MicroRNA-154: a novel candidate for diagnosis and therapy of human cancers. Onco Targets Ther. 2020;13:6603–15. https://doi.org/10.2147/OTT.S249268.
128. Kosaka N, Iguchi H, Ochiya T. Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. Cancer Sci. 2010;101:2087–92.
129. Starczynowski DT, Kuchenbauer F, Argiropoulos B, et al. Identification of miR-145 and miR-146a as mediators of the 5q-syndrome phenotype. Nat Med. 2010;16:49–58. https://doi.org/10.1038/nm.2054.
130. Wang WX, Rajeev BW, Stromberg AJ, et al. The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of β-site amyloid precursor protein-cleaving enzyme 1. J Neurosci. 2008;28:1213–23. https://doi.org/10.1523/JNEUROSCI.5065-07.2008.

131. Hébert SS, Horré K, Nicolaï L, et al. Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/β-secretase expression. Proc Natl Acad Sci U S A. 2008;105:6415–20. https://doi.org/10.1073/pnas.0710263105.
132. Boissonneault V, Plante I, Rivest S, Provost P. MicroRNA-298 and microRNA-328 regulate expression of mouse β-amyloid precursor protein-converting enzyme 1. J Biol Chem. 2009;284:1971–81. https://doi.org/10.1074/jbc.M807530200.
133. Faghihi MA, Modarresi F, Khalil AM, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. Nat Med. 2008;14:723–30. https://doi.org/10.1038/nm1784.
134. Williams AH, Valdez G, Moresi V, et al. MicroRNA-206 delays ALS progression and promotes regeneration of neuromuscular synapses in mice. Science. 2009;326:1549–54. https://doi.org/10.1126/science.1181046.
135. Nishimoto Y, Nakagawa S, Hirose T, et al. The long non-coding RNA nuclear-enriched abundant transcript 1-2 induces paraspeckle formation in the motor neuron during the early phase of amyotrophic lateral sclerosis. Mol Brain. 2013;6:31. https://doi.org/10.1186/1756-6606-6-31.
136. Gagliardi S, Zucca S, Pandini C, et al. Long non-coding and coding RNAs characterization in Peripheral Blood Mononuclear Cells and Spinal Cord from Amyotrophic Lateral Sclerosis patients. Sci Rep. 2018;8:2378. https://doi.org/10.1038/s41598-018-20679-5.
137. Yang B, Lin H, Xiao J, et al. The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2. Nat Med. 2007;13:486–91. https://doi.org/10.1038/nm1569.
138. Cordes KR, Sheehy NT, White MP, et al. MiR-145 and miR-143 regulate smooth muscle cell fate and plasticity. Nature. 2009;460:705–10. https://doi.org/10.1038/nature08195.
139. Nicoli S, Standley C, Walker P, et al. MicroRNA-mediated integration of haemodynamics and Vegf signalling during angiogenesis. Nature. 2010;464:1196–200. https://doi.org/10.1038/nature08889.
140. Fang Y, Shi C, Manduchi E, et al. MicroRNA-10a regulation of proinflammatory phenotype in athero-susceptible endothelium in vivo and in vitro. Proc Natl Acad Sci U S A. 2010;107:13450–5. https://doi.org/10.1073/pnas.1002120107.
141. Burd CE, Jeck WR, Liu Y, et al. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. PLoS Genet. 2010;6:1–15. https://doi.org/10.1371/journal.pgen.1001233.
142. Eggermann T. Silver-Russell and Beckwith-Wiedemann syndromes: opposite (epi)mutations in 11p15 result in opposite clinical pictures. Horm Res. 2009;71:30–5.
143. Fang Z, Dai W, Wang X, et al. Circulating miR-205: a promising biomarker for the detection and prognosis evaluation of bladder cancer. Tumor Biol. 2016;37:8075–82. https://doi.org/10.1007/s13277-015-4698-y.
144. Dong XY, Guo P, Boyd J, et al. Implication of snoRNA U50 in human breast cancer. J Genet Genomics. 2009;36:447–54. https://doi.org/10.1016/S1673-8527(08)60134-4.
145. Kumar S, Nag A, Mandal C. A comprehensive review on miR-200c, a promising cancer biomarker with therapeutic potential. Curr Drug Targets. 2015;16:1381–403. https://doi.org/10.2174/1389450116666150325231419.
146. Gao Y, Feng B, Han S, et al. The roles of MicroRNA-141 in human cancers: from diagnosis to treatment. Cell Physiol Biochem. 2016;38:427–48. https://doi.org/10.1159/000438641.
147. Zhang L, Xing M, Wang X, et al. MiR-148a suppresses invasion and induces apoptosis of breast cancer cells by regulating USP4 and BIM expression: e-Century Publishing Corporation; 2017.
148. Lujambio A, Portela A, Liz J, et al. CpG island hypermethylation-associated silencing of non-coding RNAs transcribed from ultraconserved regions in human cancer. Oncogene. 2010;29:6390–401. https://doi.org/10.1038/onc.2010.361.
149. Xu S, Kong D, Chen Q, et al. Oncogenic long noncoding RNA landscape in breast cancer. Mol Cancer. 2017;16:1–15. https://doi.org/10.1186/s12943-017-0696-6.

150. Thum T, Gross C, Fiedler J, et al. MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts. Nature. 2008;456:980–4. https://doi.org/10.1038/nature07511.
151. Minor J, Wang X, Zhang F, et al. Methylation of microRNA-9 is a specific and sensitive biomarker for oral and oropharyngeal squamous cell carcinomas. Oral Oncol. 2012;48:73–8. https://doi.org/10.1016/j.oraloncology.2011.11.006.
152. Chen X, He D, Da Dong X, et al. MicroRNA-124a is epigenetically regulated and acts as a tumor suppressor by controlling multiple targets in uveal melanoma. Investig Ophthalmol Vis Sci. 2013;54:2248–56. https://doi.org/10.1167/iovs.12-10977.
153. Yang Y, Huang JQ, Zhang X, Shen LF. MiR-129-2 functions as a tumor suppressor in glioma cells by targeting HMGB1 and is down-regulated by DNA methylation. Mol Cell Biochem. 2015;404:229–39. https://doi.org/10.1007/s11010-015-2382-6.
154. Mahmoudi E, Cairns MJ. MiR-137: an important player in neural development and neoplastic transformation. Mol Psychiatry. 2017;22:44–55.
155. Brest P, Lapaquette P, Souidi M, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet. 2011;43:242–5.
156. Gillen AE, Gosalia N, Leir SH, Harris A. microRNA regulation of expression of the cystic fibrosis transmembrane conductance regulator gene. Biochem J. 2011;438:25–32. https://doi.org/10.1042/BJ20110672.
157. Ramachandran S, Karp PH, Jiang P, et al. A microRNA network regulates expression and biosynthesis of wild-type and ΔF508 mutant cystic fibrosis transmembrane conductance regulator. Proc Natl Acad Sci U S A. 2012;109:13362–7. https://doi.org/10.1073/pnas.1210906109.
158. Ramachandran S, Karp PH, Osterhaus SR, et al. Post-transcriptional regulation of cystic fibrosis transmembrane conductance regulator expression and function by MicroRNAs. Am J Respir Cell Mol Biol. 2013;49:544–51. https://doi.org/10.1165/rcmb.2012-0430OC.
159. Hassan F, Nuovo GJ, Crawford M, et al. MiR-101 and miR-144 Regulate the Expression of the CFTR Chloride Channel in the Lung. PLoS One. 2012;7:e50837. https://doi.org/10.1371/journal.pone.0050837.
160. Oglesby IK, Chotirmall SH, McElvaney NG, Greene CM. Regulation of cystic fibrosis transmembrane conductance regulator by microRNA-145, -223, and -494 is altered in ΔF508 cystic fibrosis airway epithelium. J Immunol. 2013;190:3354–62. https://doi.org/10.4049/jimmunol.1202960.
161. Fabbri E, Borgatti M, Montagner G, et al. Expression of microRNA-93 and interleukin-8 during Pseudomonas aeruginosa-mediated induction of proinflammatory responses. Am J Respir Cell Mol Biol. 2014;50:1144–55. https://doi.org/10.1165/rcmb.2013-0160OC.
162. Fabbri E, Tamanini A, Jakova T, et al. A peptide nucleic acid against microrna miR-145-5p enhances the expression of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) in Calu-3 Cells. Molecules. 2018;23:71. https://doi.org/10.3390/molecules23010071.
163. Pierdomenico AM, Patruno S, Codagnone M, et al. MicroRNA-181b is increased in cystic fibrosis cells and impairs lipoxin A4 receptor-dependent mechanisms of inflammation resolution and antimicrobial defense. Sci Rep. 2017;7:129–40. https://doi.org/10.1038/s41598-017-14055-y.
164. Sonneville F, Ruffin M, Coraux C, et al. MicroRNA-9 downregulates the ANO1 chloride channel and contributes to cystic fibrosis lung pathology. Nat Commun. 2017;8:710. https://doi.org/10.1038/s41467-017-00813-z.
165. McKiernan PJ, Molloy K, Cryan SA, et al. Long noncoding RNA are aberrantly expressed in vivo in the cystic fibrosis bronchial epithelium. Int J Biochem Cell Biol. 2014;52:184–91. https://doi.org/10.1016/j.biocel.2014.02.022.
166. Saayman SM, Ackley A, Burdach J, et al. Long non-coding RNA BGas regulates the cystic fibrosis transmembrane conductance regulator. Mol Ther. 2016;24:1351–7. https://doi.org/10.1038/mt.2016.112.

167. Balloy V, Koshy R, Perra L, et al. Bronchial epithelial cells from cystic fibrosis patients express a specific long non-coding RNA signature upon Pseudomonas aeruginosa infection. Front Cell Infect Microbiol. 2017;7:218. https://doi.org/10.3389/fcimb.2017.00218.
168. Lewis MA, Quint E, Glazier AM, et al. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. Nat Genet. 2009;41:614–8. https://doi.org/10.1038/ng.369.
169. Kuhn DE, Nuovo GJ, Terry AV, et al. Erratum: chromosome 21-derived microRNAs provide an etiological basis for aberrant protein expression in human down syndrome brains (Journal of Biological Chemistry (2010) 285 (1529–1543)). J Biol Chem. 2013;288:4228.
170. Cacchiarelli D, Legnini I, Martone J, et al. miRNAs as serum biomarkers for Duchenne muscular dystrophy. EMBO Mol Med. 2011;3:258–65. https://doi.org/10.1002/emmm.201100133.
171. Hu J, Kong M, Ye Y, et al. Serum miR-206 and other muscle-specific microRNAs as non-invasive biomarkers for Duchenne muscular dystrophy. J Neurochem. 2014;129:877–83. https://doi.org/10.1111/jnc.12662.
172. Twayana S, Legnini I, Cesana M, et al. Biogenesis and function of non-coding RNAs in muscle differentiation and in Duchenne muscular dystrophy. Biochem Soc Trans. 2013;41:844–9. https://doi.org/10.1042/BST20120353.
173. Ballarino M, Cazzella V, D'Andrea D, et al. Novel long noncoding RNAs (lncRNAs) in myogenesis: a miR-31 overlapping lncRNA transcript controls myoblast differentiation. Mol Cell Biol. 2015;35:728–36. https://doi.org/10.1128/mcb.01394-14.
174. Perry MM, Muntoni F. Noncoding RNAs and Duchenne muscular dystrophy. Epigenomics. 2016;8:1527–37. https://doi.org/10.2217/epi-2016-0088.
175. Harafuji N, Schneiderat P, Walter MC, Chen YW. MiR-411 is up-regulated in FSHD myoblasts and suppresses myogenic factors. Orphanet J Rare Dis. 2013;8:55. https://doi.org/10.1186/1750-1172-8-55.
176. Vizoso M, Esteller M. The activatory long non-coding RNA DBE-T reveals the epigenetic etiology of facioscapulohumeral muscular dystrophy. Cell Res. 2012;22:1413–5. https://doi.org/10.1038/cr.2012.93.
177. Hervé M, Ibrahim EC. MicroRNA screening identifies a link between NOVA1 expression and a low level of IKAP in familial dysautonomia. DMM Dis Model Mech. 2016;9:899–909. https://doi.org/10.1242/dmm.025841.
178. Tsai KW, Hu LY, Wu CW, et al. Epigenetic regulation of miR-196b expression in gastric cancer. Genes Chromosom Cancer. 2010;49:969–80. https://doi.org/10.1002/gcc.20804.
179. Yu X, Song H, Xia T, et al. Growth inhibitory effects of three miR-129 family members on gastric cancer. Gene. 2013;532:87–93. https://doi.org/10.1016/j.gene.2013.09.048.
180. Manca S, Magrelli A, Cialfi S, et al. Oxidative stress activation of miR-125b is part of the molecular switch for Hailey-Hailey disease manifestation. Exp Dermatol. 2011;20:932–7. https://doi.org/10.1111/j.1600-0625.2011.01359.x.
181. Langevin SM, Stone RA, Bunker CH, et al. MicroRNA-137 promoter methylation is associated with poorer overall survival in patients with squamous cell carcinoma of the head and neck. Cancer. 2011;117:1454–62. https://doi.org/10.1002/cncr.25689.
182. Li D, Liu X, Lin L, et al. MicroRNA-99a inhibits hepatocellular carcinoma growth and correlates with prognosis of patients with hepatocellular carcinoma. J Biol Chem. 2011;286:36677–85. https://doi.org/10.1074/jbc.M111.270561.
183. Lin XJ, Fang JH, Yang XJ, et al. Hepatocellular carcinoma cell-secreted exosomal microRNA-210 promotes angiogenesis in vitro and in vivo. Mol Ther Nucl Acids. 2018;11:243–52. https://doi.org/10.1016/j.omtn.2018.02.014.
184. Gatto S, Della Ragione F, Cimmino A, et al. Epigenetic alteration of microRNAs in DNMT3B-mutated patients of ICF syndrome. Epigenetics. 2010;5:427–43. https://doi.org/10.4161/epi.5.5.11999.
185. Martínez F, Monfort S, Rosellá M, et al. Enrichment of ultraconserved elements among genomic imbalances causing mental delay and congenital anomalies. BMC Med Genet. 2010;3:54. https://doi.org/10.1186/1755-8794-3-54.

186. Guibinga GH. MicroRNAs: tools of mechanistic insights and biological therapeutics discovery for the rare neurogenetic syndrome Lesch-Nyhan disease (LND). Adv Genet. 2015;90:103–31. https://doi.org/10.1016/bs.adgen.2015.06.001.
187. Wojcik SE, Rossi S, Shimizu M, et al. Non-codingRNA sequence variations in human chronic lymphocytic leukemia and colorectal cancer. Carcinogenesis. 2010;31:208–15. https://doi.org/10.1093/carcin/bgp209.
188. Id Said B, Malkin D. A functional variant in miR-605 modifies the age of onset in Li-Fraumeni syndrome. Cancer Genet. 2015;208:47–51. https://doi.org/10.1016/j.cancergen.2014.12.003.
189. Mirzaei H, Gholamin S, Shahidsales S, et al. MicroRNAs as potential diagnostic and prognostic biomarkers in melanoma. Eur J Cancer. 2016;53:25–32.
190. Wang W, Zhuang Q, Ji K, et al. Identification of miRNA, lncRNA and mRNA-associated ceRNA networks and potential biomarker for MELAS with mitochondrial DNA A3243G mutation. Sci Rep. 2017b;7:1–13. https://doi.org/10.1038/srep41639.
191. Zuntini M, Salvatore M, Pedrini E, et al. MicroRNA profiling of multiple osteochondromas: identification of disease-specific and normal cartilage signatures. Clin Genet. 2010;78:507–16. https://doi.org/10.1111/j.1399-0004.2010.01490.x.
192. Gambardella S, Rinaldi F, Lepore SM, et al. Overexpression of microRNA-206 in the skeletal muscle from myotonic dystrophy type 1 patients. J Transl Med. 2010;8:48. https://doi.org/10.1186/1479-5876-8-48.
193. Fritegotto C, Ferrati C, Pegoraro V, Angelini C. Micro-RNA expression in muscle and fiber morphometry in myotonic dystrophy type 1. Neurol Sci. 2017;38:619–25. https://doi.org/10.1007/s10072-017-2811-2.
194. Wheeler TM, Leger AJ, Pandey SK, et al. Targeting nuclear RNA for in vivo correction of myotonic dystrophy. Nature. 2012;488:111–7. https://doi.org/10.1038/nature11362.
195. Li M, Lee KF, Lu Y, et al. Frequent amplification of a chr19q13.41 microRNA polycistron in aggressive primitive neuroectodermal brain tumors. Cancer Cell. 2009b;16:533–46. https://doi.org/10.1016/j.ccr.2009.10.025.
196. Kan T, Meltzer SJ. MicroRNAs in Barrett's esophagus and esophageal adenocarcinoma. Curr Opin Pharmacol. 2009;9:727–32.
197. Omura N, Li CP, Li A, et al. Genome-wide profiling of methylated promoters in pancreatic adenocarcinoma. Cancer Biol Ther. 2008;7:1146–56. https://doi.org/10.4161/cbt.7.7.6208.
198. Khan MA, Zubair H, Srivastava SK, et al. Insights into the role of micrornas in pancreatic cancer pathogenesis: potential for diagnosis, prognosis, and therapy. Adv Exp Med Biol. 2015;889:71–87.
199. Huang X, Zhi X, Gao Y, et al. LncRNAs in pancreatic cancer. Oncotarget. 2016;7:57379–90.
200. Gehrke S, Imai Y, Sokol N, Lu B. Pathogenic LRRK2 negatively regulates microRNA-mediated translational repression. Nature. 2010;466:637–41. https://doi.org/10.1038/nature09191.
201. Sahoo T, Del Gaudio D, German JR, et al. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet. 2008;40:719–21. https://doi.org/10.1038/ng.158.
202. Williamson CM, Ball ST, Dawson C, et al. Uncoupling antisense-mediated silencing and DNA methylation in the imprinted Gnas cluster. PLoS Genet. 2011;7:e1001347. https://doi.org/10.1371/journal.pgen.1001347.
203. Kim JD, Lee A, Choi J, et al. Epigenetic modulation as a therapeutic approach for pulmonary arterial hypertension. Exp Mol Med. 2015;47:e175.
204. Sun Z, Nie X, Sun S, et al. Long non-coding RNA MEG3 downregulation triggers human pulmonary artery smooth muscle cell proliferation and migration via the p53 signaling pathway. Cell Physiol Biochem. 2017;42:2569–81. https://doi.org/10.1159/000480218.
205. Chen J, Guo J, Cui X, et al. The long noncoding RNA LnRPT is regulated by PDGF-BB and modulates the proliferation of pulmonary artery smooth muscle cells. Am J Respir Cell Mol Biol. 2018;58:181–93.

206. Urdinguio RG, Fernandez AF, Lopez-Nieva P, et al. Disrupted microrna expression caused by Mecp2 loss in a mouse model of Rett syndrome. Epigenetics. 2010;5:656–63. https://doi.org/10.4161/epi.5.7.13055.
207. Wu H, Tao J, Chen PJ, et al. Genome-wide analysis reveals methyl-CpG-binding protein 2-dependent regulation of microRNAs in a mouse model of Rett syndrome. Proc Natl Acad Sci U S A. 2010;107:18161–6. https://doi.org/10.1073/pnas.1005595107.
208. Petazzi P, Sandoval J, Szczesna K, et al. Dysregulation of the long non-coding RNA transcriptome in a Rett syndrome mouse model. RNA Biol. 2013;10:1197–203. https://doi.org/10.4161/rna.24286.
209. Pauley KM, Cha S. miRNA-146a in rheumatoid arthritis: a new therapeutic strategy. Immunotherapy. 2011;3:829–31. https://doi.org/10.2217/imt.11.70.
210. Ballabio E, Mitchell T, Van Kester MS, et al. MicroRNA expression in Sézary syndrome: Identification, function, and diagnostic potential. Blood. 2010;116:1105–13. https://doi.org/10.1182/blood-2009-12-256719.
211. Qin Y, Buermans HPJ, Van Kester MS, et al. Deep-sequencing analysis reveals that the miR-199a2/214 cluster within DNM3os represents the vast majority of aberrantly expressed MicroRNAs in sézary syndrome. J Invest Dermatol. 2012;132:1520–2.
212. Haramati S, Chapnik E, Sztainberg Y, et al. miRNA malfunction causes spinal motor neuron disease. Proc Natl Acad Sci U S A. 2010;107:13111–6. https://doi.org/10.1073/pnas.1006151107.
213. Lee Y, Samaco RC, Gatchel JR, et al. miR-19, miR-101 and miR-130 co-regulate ATXN1 levels to potentially modulate SCA1 pathogenesis. Nat Neurosci. 2008;11:1137–9. https://doi.org/10.1038/nn.2183.
214. Temple IK, Shield JPH. Transient neonatal diabetes, a disorder of imprinting. J Med Genet. 2002;39:872–5.
215. Paco S, Casserras T, Rodríguez MA, et al. Transcriptome analysis of ullrich congenital muscular dystrophy fibroblasts reveals a disease extracellular matrix signature and key molecular regulators. PLoS One. 2015;10:e0145107. https://doi.org/10.1371/journal.pone.0145107.
216. Roy P, Bhattacharya G, Lahiri A, et al. Hsa-miR-503 is downregulated in β thalassemia major. Acta Haematol. 2012;128:187–9. https://doi.org/10.1159/000339492.
217. Lulli V, Romania P, Morsilli O, et al. MicroRNA-486-3p regulates γ-globin expression in human erythroid cells by directly modulating BCL11A. PLoS One. 2013;8:60436. https://doi.org/10.1371/journal.pone.0060436.
218. Siwaponanan P, Fucharoen S, Sirankapracha P, et al. Elevated levels of miR-210 correlate with anemia in β-thalassemia/HbE patients. Int J Hematol. 2016;104:338–43. https://doi.org/10.1007/s12185-016-2032-0.
219. Saki N, Abroun S, Soleimani M, et al. MicroRNA expression in β-Thalassemia and sickle cell disease: a role in the induction of fetal hemoglobin. Cell J. 2016;17:583–92.
220. Leecharoenkiat K, Tanaka Y, Harada Y, et al. Plasma microRNA-451 as a novel hemolytic marker for β0-thalassemia/HbE disease. Mol Med Rep. 2017;15:2495–502. https://doi.org/10.3892/mmr.2017.6326.
221. Srinoun K, Nopparatana C, Wongchanchailert M, Fucharoen S. MiR-155 enhances phagocytic activity of β-thalassemia/HbE monocytes via targeting of BACH1. Int J Hematol. 2017;106:638–47. https://doi.org/10.1007/s12185-017-2291-4.
222. Gasparello J, Fabbri E, Bianchi N, et al. BCL11A mRNA targeting by miR-210: a possible network regulating γ-globin gene expression. Int J Mol Sci. 2017;18:2530. https://doi.org/10.3390/ijms18122530.
223. Lai K, Jia S, Yu S, et al. Genome-wide analysis of aberrantly expressed lncRNAs and miRNAs with associated co-expression and ceRNA networks in β-thalassemia and hereditary persistence of fetal hemoglobin. Oncotarget. 2017;8:49931–43. https://doi.org/10.18632/oncotarget.18263.
224. Ma J, Liu F, Du X, et al. Changes in lncRNAs and related genes in β-thalassemia minor and β-thalassemia major. Front Med. 2017;11:74–86. https://doi.org/10.1007/s11684-017-0503-1.

225. Morrison TA, Wilcox I, Luo HY, et al. A long noncoding RNA from the HBS1L-MYB intergenic region on chr6q23 regulates human fetal hemoglobin expression. Blood Cells Mol Dis. 2018;69:1–9. https://doi.org/10.1016/j.bcmd.2017.11.003.
226. Liang C, Zhang L, Lian X, et al. Circulating exosomal SOCS2-AS1 acts as a novel biomarker in predicting the diagnosis of coronary artery disease. Biomed Res Int. 2020; https://doi.org/10.1155/2020/9182091.
227. Karakas M, Schulte C, Appelbaum S, et al. Circulating microRNAs strongly predict cardiovascular death in patients with coronary artery disease-results from the large AtheroGene study. Eur Heart J. 2017;38:516–23. https://doi.org/10.1093/eurheartj/ehw250.
228. Schulte C, Molz S, Appelbaum S, et al. miRNA-197 and miRNA-223 predict cardiovascular death in a cohort of patients with symptomatic coronary artery disease. PLoS One. 2015;10:e0145930. https://doi.org/10.1371/journal.pone.0145930.
229. Jakob P, Kacprowski T, Briand-Schumacher S, et al. Profiling and validation of circulating microRNAs for cardiovascular events in patients presenting with ST-segment elevation myocardial infarction. Eur Heart J. 2017;38:511–5. https://doi.org/10.1093/eurheartj/ehw563.
230. Zampetaki A, Willeit P, Tilling L, et al. Prospective study on circulating microRNAs and risk of myocardial infarction. J Am Coll Cardiol. 2012;60:290–9. https://doi.org/10.1016/j.jacc.2012.03.056.
231. Jansen F, Yang X, Proebsting S, et al. MicroRNA expression in circulating microvesicles predicts cardiovascular events in patients with coronary artery disease. J Am Heart Assoc. 2014;3:e001249. https://doi.org/10.1161/JAHA.114.001249.
232. Zhong L, Simard MJ, Huot J. Endothelial microRNAs regulating the NF-κB pathway and cell adhesion molecules during inflammation. FASEB J. 2018;32:4070–84. https://doi.org/10.1096/fj.201701536R.
233. Zou H, Wu LX, Tan L, et al. Significance of single-nucleotide variants in long intergenic non-protein coding RNAs. Front Cell Dev Biol. 2020;8:347.
234. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: involvement of micro-RNAs and long non-coding RNAs in disease. Biochim Biophys Acta Mol basis Dis. 2014;1842:1910–22.
235. Eichler EE, Nickerson DA, Altshuler D, et al. Completing the map of human genetic variation. Nature. 2007;447:161–5. https://doi.org/10.1038/447161a.
236. Khurana E, Fu Y, Chakravarty D, et al. Role of non-coding sequence variants in cancer. Nat Rev Genet. 2016;17:93–108.
237. Ng MCY, Graff M, Lu Y, et al. Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. PLoS Genet. 2017;13:e1006719. https://doi.org/10.1371/journal.pgen.1006719.
238. Ma G, Gu D, Lv C, et al. Genetic variant in 8q24 is associated with prognosis for gastric cancer in a Chinese population. J Gastroenterol Hepatol. 2015;30:689–95. https://doi.org/10.1111/jgh.12801.
239. Hu L, Chen SH, Lv QL, et al. Clinical significance of long non-coding RNA CASC8 rs10505477 polymorphism in lung cancer susceptibility, platinum-based chemotherapy response, and toxicity. Int J Environ Res Public Health. 2016;13:545. https://doi.org/10.3390/ijerph13060545.
240. Teerlink CC, Leongamornlert D, Dadaev T, et al. Genome-wide association of familial prostate cancer cases identifies evidence for a rare segregating haplotype at 8q24.21. Hum Genet. 2016;135:923–38. https://doi.org/10.1007/s00439-016-1690-6.
241. Kim T, Cui R, Jeon YJ, et al. Long-range interaction and correlation between MYC enhancer and oncogenic long noncoding RNA CARLo-5. Proc Natl Acad Sci U S A. 2014;111:4173–8. https://doi.org/10.1073/pnas.1400350111.
242. Zhao X, Wei X, Zhao L, et al. The rs6983267 SNP and long non-coding RNA *CARLo-5* are associated with endometrial carcinoma. Environ Mol Mutagen. 2016;57:508–15. https://doi.org/10.1002/em.22031.

243. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet. 2007;39:645–9. https://doi.org/10.1038/ng2022.
244. Tuupanen S, Turunen M, Lehtonen R, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 2009;41:885–90. https://doi.org/10.1038/ng.406.
245. Ling H, Spizzo R, Atlasi Y, et al. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. Genome Res. 2013;23:1446–61. https://doi.org/10.1101/gr.152942.112.
246. Shah MY, Ferracin M, Pileczki V, et al. Cancer-associated rs6983267 SNP and its accompanying long noncoding RNA CCAT2 induce myeloid malignancies via unique SNP-specific RNA mutations. Genome Res. 2018;28:432–47. https://doi.org/10.1101/gr.225128.117.
247. Guo H, Ahmed M, Zhang F, et al. Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. Nat Genet. 2016;48:1142–50. https://doi.org/10.1038/ng.3637.
248. Chung S, Nakagawa H, Uemura M, et al. Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. Cancer Sci. 2011;102:245–52. https://doi.org/10.1111/j.1349-7006.2010.01737.x.
249. Li L, Sun R, Liang Y, et al. Association between polymorphisms in long non-coding RNA PRNCR1 in 8q24 and risk of colorectal cancer. J Exp Clin Cancer Res. 2013a;32:104. https://doi.org/10.1186/1756-9966-32-104.
250. Zhang Z, Zhu Z, Zhang B, et al. Frequent mutation of rs13281615 and its association with PVT1 expression and cell proliferation in breast cancer. J Genet Genomics. 2014;41:187–95. https://doi.org/10.1016/j.jgg.2014.03.006.
251. Hanson RL, Craig DW, Millis MP, et al. Identification of PVT1 as a candidate gene for end-stage renal disease in type 2 diabetes using a pooling-based genome-wide single nucleotide polymorphism association study. Diabetes. 2007;56:975–83. https://doi.org/10.2337/db06-1072.
252. Meyer KB, Maia AT, O'Reilly M, et al. A functional variant at a prostate cancer predisposition locus at 8q24 is associated with PVT1 expression. PLoS Genet. 2011;7:e1002165. https://doi.org/10.1371/journal.pgen.1002165.
253. Cerhan JR, Berndt SI, Vijai J, et al. Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. Nat Genet. 2014;46:1233–8. https://doi.org/10.1038/ng.3105.
254. Zuo X, Wang H, Mi Y, et al. The association of casc16 variants with breast cancer risk in a northwest Chinese female population. Mol Med. 2020;26:11. https://doi.org/10.1186/s10020-020-0137-7.
255. Maris JM, Mosse YP, Bradfield JP, et al. Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. N Engl J Med. 2008;358:2585–93. https://doi.org/10.1056/NEJMoa0708698.
256. Tao R, Hu S, Wang S, et al. Association between indel polymorphism in the promoter region of lncRNA GAS5 and the risk of hepatocellular carcinoma. Carcinogenesis. 2015;36:1136–43. https://doi.org/10.1093/carcin/bgv099.
257. Li Q, Ma G, Sun S, et al. Polymorphism in the promoter region of lncRNA GAS5 is functionally associated with the risk of gastric cancer. Clin Res Hepatol Gastroenterol. 2018a;42:478–82. https://doi.org/10.1016/j.clinre.2018.01.006.
258. Gao W, Zhu M, Wang H, et al. Association of polymorphisms in long non-coding RNA H19 with coronary artery disease risk in a Chinese population. Mutat Res - Fundam Mol Mech Mutagen. 2015;772:15–22. https://doi.org/10.1016/j.mrfmmm.2014.12.009.
259. Wu Q, Yan W, Han R, et al. Polymorphisms in long noncoding RNA H19 contribute to the protective effects of coal workers' pneumoconiosis in a Chinese population. Int J Environ Res Public Health. 2016;13:903. https://doi.org/10.3390/ijerph13090903.

260. Chu M, Yuan W, Wu S, et al. Quantitative assessment of polymorphisms in H19 lncRNA and cancer risk: a meta-analysis of 13,392 cases and 18,893 controls. Oncotarget. 2016;7:78631–9. https://doi.org/10.18632/oncotarget.12530
261. Verhaegh GW, Verkleij L, Vermeulen SHHM, et al. Polymorphisms in the H19 gene and the risk of bladder cancer. Eur Urol. 2008;54:1118–26. https://doi.org/10.1016/j.eururo.2008.01.060.
262. Yang ML, Huang Z, Wang Q, et al. The association of polymorphisms in lncRNA-H19 with hepatocellular cancer risk and prognosis. Biosci Rep. 2018a;38:BSR20171652. https://doi.org/10.1042/BSR20171652.
263. Wang BG, Lv Z, Ding HX, et al. The association of lncRNA-HULC polymorphisms with hepatocellular cancer risk and prognosis. Gene. 2018a;670:148–54. https://doi.org/10.1016/j.gene.2018.05.096.
264. Zheng J, Huang X, Tan W, et al. Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with PTPN11 degradation. Nat Genet. 2016;48:747–57. https://doi.org/10.1038/ng.3568.
265. Wu H, Zheng J, Deng J, et al. A genetic polymorphism in lincRNA-uc003opf.1 is associated with susceptibility to esophageal squamous cell carcinoma in Chinese populations. Carcinogenesis. 2013;34:2908–17. https://doi.org/10.1093/carcin/bgt252.
266. Fava VM, Manry J, Cobat A, et al. A genome wide association study identifies a lncRna as risk factor for pathological inflammatory responses in leprosy. PLoS Genet. 2017;13:e1006637. https://doi.org/10.1371/journal.pgen.1006637.
267. Zhuo Y, Zeng Q, Zhang P, et al. Functional polymorphism of lncRNA MALAT1 contributes to pulmonary arterial hypertension susceptibility in Chinese people. Clin Chem Lab Med. 2017;55:38–46. https://doi.org/10.1515/cclm-2016-0056.
268. Li Q, Zhu W, Zhang B, et al. The MALAT1 gene polymorphism and its relationship with the onset of congenital heart disease in Chinese. Biosci Rep. 2018b;38:BSR20171381. https://doi.org/10.1042/BSR20171381.
269. Li Y, Bao C, Gu S, et al. Associations between novel genetic variants in the promoter region of MALAT1 and risk of colorectal cancer. Oncotarget. 2017b;8:92604–14. https://doi.org/10.18632/oncotarget.21507.
270. Wang BG, Xu Q, Lv Z, et al. Association of twelve polymorphisms in three onco-lncRNa genes with hepatocellular cancer risk and prognosis: a case-control study. World J Gastroenterol. 2018b;24:2482–90. https://doi.org/10.3748/wjg.v24.i23.2482.
271. Wallace C, Smyth DJ, Maisuria-Armer M, et al. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat Genet. 2010;42:68–71. https://doi.org/10.1038/ng.493.
272. Westra HJ, Martínez-Bonet M, Onengut-Gumuscu S, et al. Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. Nat Genet. 2018;50:1366–74.
273. Ishii N, Ozaki K, Sato H, et al. Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J Hum Genet. 2006;51:1087–99. https://doi.org/10.1007/s10038-006-0070-9.
274. Rao SQ, Hu HL, Ye N, et al. Genetic variants in long non-coding RNA MIAT contribute to risk of paranoid schizophrenia in a Chinese Han population. Schizophr Res. 2014a;166:125–30. https://doi.org/10.1016/j.schres.2015.04.032.
275. Xue Y, Wang M, Kang M, et al. Association between lncrna PCGEM1 polymorphisms and prostate cancer risk. Prostate Cancer Prostatic Dis. 2013;16:139–44. https://doi.org/10.1038/pcan.2013.6.
276. Gao P, Xia JH, Sipeky C, et al. Biology and clinical implications of the 19q13 aggressive prostate cancer susceptibility locus. Cell. 2018;174:576–589.e18. https://doi.org/10.1016/j.cell.2018.06.003.
277. He H, Li W, Liyanarachchi S, et al. Genetic predisposition to papillary thyroid carcinoma: Involvement of FOXE1, TSHR, and a novel lincRNA Gene, PTCSC2. J Clin Endocrinol Metab. 2015;100:E164–72. https://doi.org/10.1210/jc.2014-2147.

278. Jendrzejewski J, He H, Radomska HS, et al. The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proc Natl Acad Sci U S A. 2012;109:8646–51. https://doi.org/10.1073/pnas.1205654109.
279. Lee TH, Ko TM, Chen CH, et al. Identification of PTCSC3 as a novel locus for large-vessel ischemic stroke: a genome-wide association study. J Am Heart Assoc. 2015;5:e003003. https://doi.org/10.1161/JAHA.115.003003.
280. Han L, Liu S, Liang J, et al. A genetic polymorphism at miR-526b binding-site in the lincRNA-NR_024015 exon confers risk of esophageal squamous cell carcinoma in a population of North China. Mol Carcinog. 2017;56:960–71. https://doi.org/10.1002/mc.22549.
281. Zheng Y, Yang C, Tong S, et al. Genetic variation of long non-coding RNA TINCR contribute to the susceptibility and progression of colorectal cancer. Oncotarget. 2017;8:33536–43. https://doi.org/10.18632/oncotarget.16538.
282. Ge L, Wang Q, Hu S, Yang X. Rs217727 polymorphism in H19 promotes cell apoptosis by regulating the expressions of H19 and the activation of its downstream signaling pathway. J Cell Physiol. 2019;234:7279–91. https://doi.org/10.1002/jcp.27485.
283. Kulakova O, Bashinskaya V, Kiselev I, et al. Pharmacogenetics of glatiramer acetate therapy for multiple sclerosis: the impact of genome-wide association studies identified disease risk loci. Pharmacogenomics. 2017;18:1563–74. https://doi.org/10.2217/pgs-2017-0058.
284. Ghesquières H, Larrabee BR, Casasnovas O, et al. A susceptibility locus for classical Hodgkin lymphoma at 8q24 near MYC/PVT1 predicts patient outcome in two independent cohorts. Br J Haematol. 2018;180:286–90.
285. Wang J-Z, Xiang J-J, Wu L-G, et al. A genetic variant in long non-coding RNA MALAT1 associated with survival outcome among patients with advanced lung adenocarcinoma: a survival cohort analysis. BMC Cancer. 2017a;17:167. https://doi.org/10.1186/s12885-017-3151-6.
286. Gong WJ, Peng JB, Yin JY, et al. Association between well-characterized lung cancer lncRNA polymorphisms and platinum-based chemotherapy toxicity in Chinese patients with lung cancer. Acta Pharmacol Sin. 2017;38:581–90. https://doi.org/10.1038/aps.2016.164.
287. Yan H, Zhang DY, Li X, et al. Long non-coding RNA GAS5 polymorphism predicts a poor prognosis of acute myeloid leukemia in Chinese patients via affecting hematopoietic reconstitution. Leuk Lymphoma. 2017;58:1948–57. https://doi.org/10.1080/10428194.2016.1266626.
288. Guo Z, Wang Y, Zhao Y, et al. Genetic polymorphisms of long non-coding RNA GAS5 predict platinum-based concurrent chemoradiotherapy response in nasopharyngeal carcinoma patients. Oncotarget. 2017;8:62286–97. https://doi.org/10.18632/oncotarget.19725.
289. Huang D, Yi X, Zhang S, et al. GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits. Nucleic Acids Res. 2018;46:W114–20. https://doi.org/10.1093/nar/gky407.
290. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
291. Buniello A, Macarthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12. https://doi.org/10.1093/nar/gky1120.
292. Sherry ST, Ward MH, Kholodov M, et al. DbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11. https://doi.org/10.1093/nar/29.1.308.
293. Miao YR, Liu W, Zhang Q, Guo AY. LncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. Nucleic Acids Res. 2018;46:D276–80. https://doi.org/10.1093/nar/gkx1004.
294. Chen X, Hao Y, Cui Y, et al. LncVar: a database of genetic variation associated with long non-coding genes. Bioinformatics. 2017;33:112–8. https://doi.org/10.1093/bioinformatics/btw581.
295. Luo H, Bu D, Shao L, et al (2020) Single-cell long non-coding RNA landscape of T cells in human cancer immunity. bioRxiv 2020.07.22.215855. https://doi.org/10.1101/2020.07.22.215855.

296. Liu SJ, Nowakowski TJ, Pollen AA, et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. 2016;17:1–17. https://doi.org/10.1186/s13059-016-0932-1.
297. Metzker ML. Sequencing technologies the next generation. Nat Rev Genet. 2010;11:31–46.
298. Liu N, Parisien M, Dai Q, et al. Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. RNA. 2013;19:1848–56. https://doi.org/10.1261/rna.041178.113.
299. Picardi E, D'Erchia AM, Gallo A, et al. Uncovering RNA editing sites in long non-coding RNAs. Front Bioeng Biotechnol. 2014;2:64. https://doi.org/10.3389/fbioe.2014.00064.
300. Dai D, Wang H, Zhu L, et al. N6-methyladenosine links RNA metabolism to cancer progression review-article. Cell Death Dis. 2018;9:124.
301. Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct rnA sequencing on an array of nanopores. Nat Methods. 2018;15(3):201–6. https://doi.org/10.1038/Nmeth.4577.
302. Pipkin ME, Lichtenheld MG. A reliable method to display authentic DNase I hypersensitive sites at long-ranges in single-copy genes from large genomes. Nucleic Acids Res. 2006;34:e34. https://doi.org/10.1093/nar/gkl006.
303. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 2015;2015:21.29.1–9. https://doi.org/10.1002/0471142727.mb2129s109.
304. Rao SSP, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014b;159:1665–80. https://doi.org/10.1016/j.cell.2014.11.021.
305. Bell JC, Jukam D, Teran NA, et al. Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. Elife. 2018;7:e27024. https://doi.org/10.7554/eLife.27024.
306. Li X, Zhou B, Chen L, et al. GriD-seq reveals the global rNA-chromatin interactome. Nat Publ Gr. 2017a;35 https://doi.org/10.1038/nbt.3968.
307. Bassett AR, Akhtar A, Barlow DP, et al. Considerations when investigating lncRNA function in vivo. Elife. 2014;3:1–14. https://doi.org/10.7554/eLife.03058.
308. Liu SJ, Horlbeck MA, Cho SW, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science. 2017;355:aah7111. https://doi.org/10.1126/science.aah7111.
309. Dinger ME, Pang KC, Mercer TR, Mattick JS. differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol. 2008;4:e1000176. https://doi.org/10.1371/journal.pcbi.1000176.
310. Slavoff SA, Mitchell AJ, Schwaid AG, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol. 2013;9:59–64. https://doi.org/10.1038/nchembio.1120.
311. Saghatelian A, Couso JP. Discovery and characterization of smORF-encoded bioactive polypeptides. Nat Chem Biol. 2015;11:909–16.
312. Amin N, McGrath A, Chen Y-PP. Evaluation of deep learning in non-coding RNA classification. Nat Mach Intell. 2019;1:246–56. https://doi.org/10.1038/s42256-019-0051-2.
313. Mackowiak SD, Zauber H, Bielow C, et al. Extensive identification and analysis of conserved small ORFs in animals. Genome Biol. 2015;16:179. https://doi.org/10.1186/s13059-015-0742-x.
314. Hung T, Wang Y, Lin MF, et al. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet. 2011;43(7):621–9. https://doi.org/10.1038/ng.848.
315. Redon S, Reichenbach P, Lingner J. The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. Nucleic Acids Res. 2010;38(17):5797–806. https://doi.org/10.1093/nar/gkq296.

Chapter 6
Satellite and Tandem DNA Repeats in the Human Genome

Luciana Amaral Haddad

6.1 Introduction

The complementarity between the two strands of the DNA molecule allows them to reassociate if temperature drops down after full denaturation by heat. By the early 1960s, the sizes of certain viral and bacterial DNA had been estimated, and reassociation data available for these genomes led to the expectation that denatured DNA of vertebrate genomes would take much longer to reassociate due to their larger sizes. It was already known that the rate of DNA single strand reassociation directly depended on the DNA concentration and number of collisions between strands, what could be experimentally controlled by immobilization of one of the two DNA strands (e.g., on cellulose, agar, nitrocellulose) and the dilution factor. Comparative quantification of single-stranded DNA could be obtained by spectrophotometry, as denatured DNA absorbs more ultraviolet light than renatured DNA does, because the nitrogenous bases become more exposed, a property known as hyperchromicity. Hence, in denaturation-reassociation experiments, plotting the product of single-stranded DNA concentration and time (after denaturation) yielded curves of DNA reassociation kinetics. Depending on the DNA sequence characteristics, the curves could differ in a few orders of magnitude (Fig. 6.1a). It is consequently possible to estimate the rate of DNA that remains denatured (C/C_o ratio, where C is molar concentration and C_o initial molar concentration). Since the time of incubation to reach, for instance, 50% of single-stranded DNA reassociation reflects the factors presented above, the product of initial DNA concentration and time (in seconds) has been a general reference for reassociation kinetics experiments, known as the C_ot value or C-value, which is specific for each species.

L. A. Haddad (✉)
Department of Genetics and Evolutionary Biology, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil
e-mail: haddadL@usp.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_6

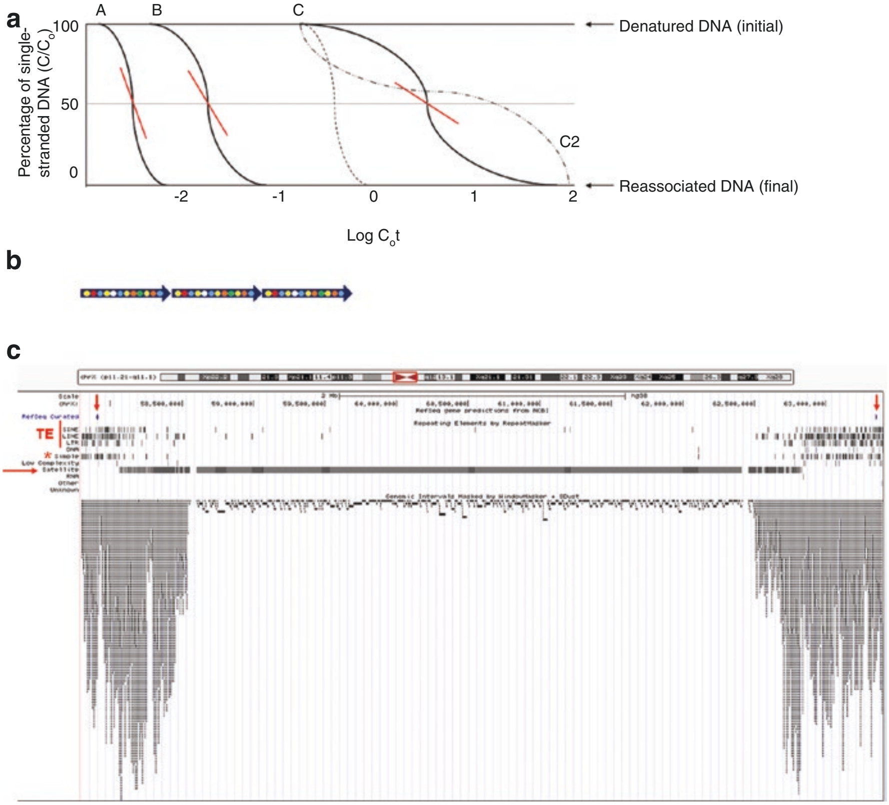


Fig. 6.1 (**a**) A scheme of a putative plot of denatured DNA reassociation kinetics, depicting three illustrative curves A, B and C, according to the relation between C$_0$t (C$_o$ X time after denaturation) and C/C$_o$ (percentage of reassociated single-stranded DNA). The fastest reassociating DNA is indicated in curve A, which could represent repetitive DNA, such as *in vitro* synthetized poly-uridine and poly-adenine. Curve B is representative of DNA of virus or bacteria genomes, whereas curve C is more characteristic of eukaryotic genomes. The broad distribution of curve C represents the summation of various curves of distinct DNA fragments, including highly repetitive DNA (satellite DNA) represented in curve C1 and unique sequences or miscellaneous non-satellite DNA illustrated in curve C2. (**b**) Three identical HORs are illustrated in tandem. In this case, each one (dark blue arrow) contains 12 alpha satellite monomers (171 bp in length each) in tandem represented by small circles. Among them, there are base substitutions (different colors). Although monomers differ within a HOR, HORs in the same centromere tend to be homogeneous. HOR repeats in tandem hundreds to thousands times. (**c**) Partial map of human X chromosome spanning 5.6 Mbp (ChrX:57,800,001–63,400,000) of the centromere, depicting the location of satellite DNA (horizontal arrow), transposable elements (TE), and simple repeats (asterisk). These repetitive sequences are indicated on the map as grey lines or blocks, depending on their lengths. The centromere is a region that lacks genes. The first genes on the short and long arms of the X chromosome are indicated by the vertical arrows on the top, left and right sides, respectively. The bottom part of the figure has the repeat elements masking indicated by bars. The central area of the figure with a long satellite line corresponds to the HOR array of alpha satellite DNA (nearly 3.1 Mbp) which is poor in other kinds of repeats. Data obtained at the University of California in Santa Cruz (UCSC) genome browser accessed in December, 2020

As the C-value was expected to be a parameter of the size of the genome, it was surprising to observe in the mid 1960s that denatured, sheared DNA of eukaryotic species submitted to reassociation analysis produced at least two types of $C_o t$ curves. One curve, as anticipated, corresponded to the much higher $C_o t$ values expected for eukaryotic genomes when compared to the genomes of viruses or bacteria, reflecting their larger sizes. By contrast, an additional curve of low $C_o t$ values was observed for a portion of DNA fragments of eukaryotic species. This unexpected evidence was reproducible for different eukaryotic organisms, and became known as 'the C-value paradox'. Fragments of DNA in solution with faster reassociation rates suggested they contained repetitive sequences that would more easily base pair [5]. At that time, DNA strand immobilization was a fundamental experimental control to rule out the possibility that circular DNA or self-complementary DNA could justify the faster reassociation curve. Consequently, the direct relationship between the C-value and genome size would apply only in the lack of repetitive sequences. Whereas the curve for pure repetitive DNA sequences has a steep slope, a broader reassociation curve with a lower general slope (Fig. 6.1a, red line) is indicative of a heterogeneous genome, comprised of a high content of repetitive sequences (low complexity sequences; Fig. 6.1a; curve C1) in addition to unique (complex) sequences (Fig. 6.1a; curve C2). The suggestion that repetitive DNA could account for the fast DNA strand reassociation rates solved the C-value paradox, and indicated for the first time that the eukaryotic genome size could probably not be proportional to its gene content (see Chaps. 1 and 4).

In the same period, cesium chloride gradient centrifugation assays disclosed a minor band that segregated distinctly from the bulk of fractionated mouse genomic DNA due to differences in buoyant densities. The minor band, termed satellite DNA, when submitted to denaturation-reassociation assays presented fast reassociation kinetics strongly suggestive of repetitive DNA [5].

6.2 Satellite DNA

Human satellite DNA was characterized as the major component of centromeric and pericentromeric chromosomal regions (including acromeric satellite repeats that occur specifically in the short arm of acrocentric chromosomes) as well as a great portion of the Y chromosome, altogether representing the largest constitutive heterochromatin blocks of the human genome (See Chaps. 1 and 2). DNA sequencing disclosed that satellite DNA is composed of long segments of tandemly repeated near-identical sequences (repeat units directly adjacent to each other) comprising a few Megabase pairs (Mbp) per chromosome. Overall, satellite repeats in heterochromatic regions should comprehend more than 5% of the human genome length (see below and Chap. 1).

At the nucleotide level, human satellite DNA is known as a diverse family of repeat sequences, with each individual member unit harboring in average more than 100 base pairs (bp), classified as medium to long pattern of tandem repeats (Open database of transposable elements and DNA repetitive families; https://dfam.org/), although diverse unit lengths are observed. Among the most common human satellite DNA family members (e.g., alpha, beta, gamma satellites, human satellites (HSAT) 1, 2 and 3, ACRO1, etc.), alpha satellite is believed to account for more than 50% of the human satellite DNA content and has been extensively employed as an experimental genomic model of satellite DNA domain organization. The alpha satellite DNA subfamily is defined as a group of related though highly divergent repeats rich in adenine (A) and thymine (T), found in every normal human centromere; having each unit (monomer) approximately 171 bp in length. Early physical maps of the organization of centromeric DNA assigned satellite DNA to fragments encompassing approximately 3 Mbp per human chromosome centromere. These observations led to content estimates of nearly 2.3% (23 chromosomes per genome X 3 Mbp/3000 Mbp of genome length) of the human genome, in agreement with the current observation of 2.4% of the latest assembly of the human genome reference (hg38) containing the 22 autosomes and both sex chromosomes, X and Y.

In human centromeres, alpha satellite DNA is found organized as tandem monomers combined to higher-order repeat (HOR) units. Alpha satellite units within tandem monomers frequently show DNA variants that make their sequences more easily individualized, thus amenable to alignment to the human genome reference. Tandem monomers may be interspersed by transposable elements, other types of satellite DNA (e.g., HSA 1 or delta satellite) or simple DNA repeats (see Sect. 6.3). By contrast, HOR typically occurs as long (Mbp-sized) homogenized arrays composed by near-identical units. Each HOR unit is defined on a chromosome-specific basis, composed of a certain number of alpha satellite monomers (e.g., 12 monomers on the X-chromosome HOR unit). As there are hundreds to thousands HOR units per centromere and they show nearly 100% identity among them, this tandemly repeated DNA array is considerably homogenous. The HOR array is flanked by degenerate tandem alpha satellite DNA monomers separating it from the chromosome arms. HOR arrays may be interrupted by other repetitive elements, such as transposable elements (Fig. 6.1c).

HSATs 2 and 3 are the most common sequence types of satellite DNA in pericentromeric regions of human chromosomes 1, 9, 16 and Y, extending over 10 Mbp in each one and showing length polymorphisms between individuals. They account for most of the human Y chromosome sequence. Although HSATs 2/3 lack array sequence consensus, they are enriched in repeats of a short, pentameric (CATTC) motif in complex arrangements. HSATs 2/3 appear to organize in subfamilies with different unit lengths specific for distinct chromosome pericentromeric regions. Although alpha satellite short reads obtained by human genome sequencing have been more reliably aligned to the human genome reference, HSATs and other classes of human satellite DNA still have limited mapping in the reference assembly. For instance, HSATs 2/3 sequences have been identified in an average frequency of 2% of the human genome by genotyping different populations, but in the current genome assembly (hg38) they correspond to 0.01% of its content.

It is known that satellite DNA is the major sequence component needed to fill in the gaps observed upon assembly of human genome sequences. These gaps corresponding to the unassembled parts of the full human genome sequence owe to the high homogeneity of thousands of tandem repeats of satellite DNA and other types of tandem repeats (see Sect. 6.3). DNA variants observed in distinct satellite repeat units would ideally aid to direct the assembly of sequence reads and align to specific chromosomal regions. Modern sequencing technologies permit to approach a more efficient assembly of centromeric and pericentromeric DNA by relying on DNA sequencing that yields long reads (up to few tens of kilobases compared to few hundreds bases of short reads), such as those developed by Pacific Biosciences (Menlo Park, CA, USA) or Oxford Nanopore (United Kingdom) technologies. However, longer reads associate with reduced coverage (number of times a base position is sequenced), thus limiting the advantage of using DNA variants for reference orientation. Therefore, linear assembly of long DNA stretches of tandemly repeated sequences is yet an unresolved challenging task that impairs identification of variants causing centromere dysfunction. Thus, few Mbp-long centromere alterations have been mostly assessed by cytogenomics exams as fluorescence *in situ* hybridization (FISH) or chromosome microarray (see Chap. 2), but the field still lacks an appropriate approach to study how nucleotide variants in satellite DNA may affect centromere assembly and chromosome segregation, which when deficient may be implicated in chromosomal instability, aneuploidy, aging and cancer.

It is highly expected that the next DNA sequencing technological breakthrough producing even longer reads, spanning hundreds of kilobases (kb), will provide fully assembled, telomere-to-telomere human chromosome sequences useful for satellite DNA functional association studies in population and medical genetics [18]. In a recent research [1], the X-chromosome sequence assembly was submitted to manual finishing of 29 reference gaps, fully resolving 1,147,861 bp of previously ambiguous bases, with nearly 99.9% accuracy and uniform mapping coverage across the entire chromosome. The canonical HOR unit of the haploid centromeric satellite array on the X chromosome (DXZ1) has approximately 2 kb in length, and is composed of 12 divergent alpha satellite monomers ordered in a head-to-tail manner. The group has tiled long reads across the entire centromeric satellite array based on catalogued structural and single-nucleotide variants within the canonical DXZ1 HOR unit. Previous assessments among X-chromosomes had shown that DXZ1 HOR array spans 2.2–3.7 Mbp (mean of 3,010 kb) with limited nucleotide differences between repeat copies. The recent data disclosed that among 7,316 DXZ1-containing high fidelity DNA long reads, 99.85% had pure satellite DNA, whereas few reads showed evidence for a transition from DXZ1 into a single insertion of a previously reported sequence derived from a long interspersed L1 element (see Chap. 8). Database analyses disclosed that among 38,875 sequenced HOR units, 98.2% had full-length canonical 12-mer repeats and 1.8% consisted of variant repeat structure with different numbers of monomers [1]. These data confirm the homogeneous nature of satellite repeat array in the X-chromosome centromere (Fig. 6.1b, c).

Centromeric satellite DNA can vary extensively among distinct chromosomes. If few specific variants in satellite DNA unit composes a unique sequence of at least few tens of bases, it can be employed in genome sequence database searches (e.g., sequence read archive (SRA) at the National Center for Biotechnology Information, Bethesda, MD, USA; https://www.ncbi.nlm.nih.gov/sra). Hence, in human population datasets of genome sequences, the frequency and extent of centromeric or pericentromeric satellite could be estimated. Based on unique oligonucleotide sequence searches, the extent of alpha satellite has been shown to vary from 1% to 5% of the human genome, whereas HSATs 2/3, although generally expected to be less abundant than the former, varied widely between 1% and 7% of the human genome length. Therefore, it is expected that constitutive heterochromatic sequences may in average consist of at least 8% of the human genome (See Chap. 1), and this rate should be even higher and variable as more information on the large diversity of satellite DNA length polymorphisms is likely to be gained in a near future.

Finally, it is important to highlight that satellite DNA repeats in the human genome classify as highly repetitive sequences because more than a thousand tandem repeats can be found in a specific locus. As mentioned, their estimated total length in the human genome (~8%) is lower than the content of transposable elements, which comprise approximately 45% of the human genome (see Chaps. 1 and 8). However, a given transposable element is longer than each satellite DNA family member unit, and individually intersperses unique sequences in the genome. Hence, the total number of their units is lesser than satellite DNA repeats and, consequently, transposable elements are considered moderately repetitive sequences.

6.3 Tandem DNA Repeats

Different loci in the human genome contain tandem DNA repeats in numbers significantly lower than satellite DNA repeats. The classification of this distinct class of DNA repeats is based on the repeat unit size and locus array length. By analogy to the satellite tandem DNA repeats, they have been termed minisatellites and microsatellites. Minisatellites can have repeat units with 7 bp to nearly 100 bp, spanning segments of 0.5 kb to several kilobases in length. Microsatellites have repeat units with 1–6 bp, and generally extend to few tens or hundreds base pairs. It is commonly seen in the literature the classification of the 6-bp motif as either micro- or minisatellite. Microsatellite's short units characterize them as simple (low complexity) sequences, simple sequence repeats (SSR) or short tandem repeats (STR), and they can have all sorts of base combination motifs though with higher frequency towards a few sequences for each motif length (e.g. $(CA)_n$ more common than $(CG)_n$ for dinucleotide microsatellites). Minisatellites are rich in cytosine (C) and guanine (G) and concentrate in subtelomeric regions, and microsatellite loci disperse in the whole genome. Both fractionate in cesium chloride gradient centrifugation together with the bulk of unique sequence DNA, and do not concentrate in constitutive heterochromatic regions. As these repetitive loci have more limited extension when compared to satellite DNA, they classify as moderately repetitive loci in the human genome. Simple sequences comprehend nearly 3% of the human genome sequence length (see Chap. 1).

Because the repeat number of micro- and minisatellite locus motifs can vary among individuals of a population and each allele is inherited in a Mendelian pattern, these loci have been classified as length/size polymorphisms. Therefore, minisatellites are commonly referred to as variable number tandem repeats (VNTR). Mini- and microsatellite length variability has been associated with transcription and splicing quantitative differences depending on the localization of the repeats, i.e., near promoter and enhancers or introns, respectively.

Tandem DNA repeats with size polymorphisms and high heterozygosity rates have been the most widely used DNA markers in forensic analysis along with mitochondrial DNA (see Chap. 10). Two general approaches have been initially employed: first, Southern blotting with multi-locus probes composed of common oligonucleotide sequence shared among different minisatellite loci (DNA fingerprinting) or with single-locus probes to genotyping two alleles from a unique locus; secondly, polymerase chain reaction (PCR) *in vitro* amplification of microsatellite loci with primers annealing to flanking unique DNA sequences. The latter approach has been also vastly employed in establishing chromosomal haplotype maps of linked highly polymorphic loci for genome projects, as well as in genetic analyses of large genealogies aiming at positional cloning of genes involved in genetic disorders based on indirect molecular identification of mutation carriers.

Today, in spite of the recognized importance of VNTR and STR as markers in population and medical genetics, their repetitive nature and, in particular, VNTR extension and relatively high C-plus-G content, can make them difficult to align to the reference genome notably when short-read sequencing is employed. Consequently, there is an underrepresentation of VNTRs in the human reference genome and a trend to present their shortest allele, limiting the genome reference as a resource for identifying potential polymorphic loci, genotyping tandem repeat loci or being a repository of the major allele for all loci (see Chap. 3). As described for the human X chromosome centromeric satellite DNA (see Sect. 6.2), human genome sequencing based on longer reads has been disclosing novel STR and VNTR loci with potential to associate with diseases (see Sect. 6.5) [8, 28].

Recent human genome sequencing producing long DNA reads confirmed the enrichment of VNTRs in the end of chromosomes, in particular to the most telomeric 5-Mbp segment. Whereas in average 55% of human minisatellites are subtelomeric, in certain chromosomes, as in the short arm of human chromosome 17, up to 85% of VNTRs have this location. Subtelomeric minisatellites tend to be richer in C and G and near genes, more often within introns, in contrast to others that disperse in the genome and appear to have originated by retrotransposition of short interspersed nuclear element-VNTR-Alu (SVA; see Chap. 8) with a bias against genes. Subtelomeric accumulation of minisatellites are likely a result of increased double-strand breakage and male meiotic recombination in the region [28].

Different VNTRs have been implicated in increased risk for human diseases, although recent association studies have mostly employed single nucleotide variants. VNTRs have been directly associated with disease etiology, as expanded alleles in homozygosity, which can repress gene transcription (see Sect. 6.4.1).

6.3.1 Microsatellite DNA (STR)

Microsatellite length variability depends on intrinsic and extrinsic factors that define for each locus a mutation rate, estimated between 1×10^{-5} to 1×10^{-3}. Microsatellites with shorter motifs (e.g., dinucleotide vs. tri- or tetranucleotide), longer extensions (several repeats) and no sequence substitutions (pure tracts) are more susceptible to repeat number alterations, creating novel alleles, with a trend to repeat gain and alteration of a single unit (±1 repeat). They thus display the highest heterozygosity rates.

Tandem DNA repeat stability can be affected by processes that lead to transient separation of DNA complementary strands, such as DNA replication or transcription. Single-stranded tandemly repeated DNA has propensity to form hairpins and slipped structures requiring efficient repair systems at common-length repeats for critically maintaining their stability. Human cells possess a DNA hairpin repair system for correct removal of repeat-produced hairpins, which appears similar to the nucleotide-excision repair pathway. Newly replicated, nicked DNA strand is mostly targeted for repair, and hairpins are removed by dual incisions or a combination of incision and endonucleolytic cleavage. During transcription, trinucleotide and tetranucleotide repeats have the propensity to form stable DNA-RNA hybrids between the DNA template and nascent RNA strand, rendering non-template DNA strand persistently unpaired increasing the odds for formation of intrastrand non-canonical DNA structures, which can be recognized by components of the mismatch repair complex (MMR). Persistent DNA-RNA hybrids and non-canonical DNA structures may lead to error-prone repair and, consequently, repeat instability.

Human proteins of the MMR system are named according to the orthologous genes of yeast: the mutL homologue 1 (MLH1), mutS homologues 2 (MSH2) and 6 (MSH6), and the Pms1 homologue 2 (PMS2). MMR deficiency is observed in certain cancer types, mainly in colorectal cancer and endometrial cancer. Loss of function of MMR protein due to pathogenic DNA variants in the respective gene results in increased tumor mutation burden, secondary to lack of DNA repair. The deficiency in MMR repair also associates with instability of microsatellites since these loci are more susceptible to replication errors. Therefore, microsatellite instability is a marker of MMR inactivity useful in analysis of tumor samples by PCR genotyping. Germline pathogenic DNA variants inactivating genes encoding any MMR protein cause Lynch syndrome that increases the risk of developing various types of cancer, in particular colorectal cancer, thus also named hereditary nonpolyposis colorectal cancer (HNPCC). Somatic mutations inactivating MMR genes are also seen in nearly 15% of non-familial colorectal cancers. The high frequency of pathogenic DNA variants in cancer resultant from deficient MMR system increases the expression of protein neoantigens, for instance, by splicing variants creating new protein isoforms or loss of translation stop codons producing translation frame reading-through or CDS fusion. Therefore, neoplasia identified with microsatellite instability (MMR deficiency) has been considered as targets for immunotherapy in clinical trials. Immune checkpoint inhibitors Pembrolizumab (KEYTRUDA®, MSD, Kenilworth, NJ, USA) and Nivolumab (OPDIVO®, Bristol Myers Squibb, New Brunswick, NJ, USA) targeting programmed cell death protein 1 (PD1) and Ipilimumab (YERVOY®, Bristol Myers Squibb) targeting cytotoxic T lymphocyte-associated protein 4 (CTLA4) have been indeed approved in several countries to treat colorectal cancer and other neoplasia with microsatellite instability [16].

6.4 Dynamic Mutations: Expansion of Transcribed, Unstable Tandem Repeats

Polymorphic microsatellites located in protein-coding genes may be found in exons or introns. Exonic microsatellites can be part of the coding sequence (CDS) in general as trinucleotides coding for one amino acid, or in the 5′ or 3′ untranslated region (UTR). When transcribed, exonic microsatellites will be part of the mature mRNA. Pathogenic unstable microsatellite alleles have been described in genes associated with genetic diseases. Their instability associates with longer tracts of repeats increasing the likelihood of intergenerational expansion into even longer alleles. The pathogenicity of novel alleles of expanded unstable microsatellites depends on their location in the gene sequence. Hence, transcribed unstable alleles have been classified according to their localization in the 5′ UTR, the CDS, the 3′ UTR or in introns. Intronic microsatellites, although transcribed are spliced out, thus not present in the mature mRNA. Transcribed unstable microsatellites have mostly trinucleotide motifs. A second tier of microsatellite classification adopts its motif sequence, having CGG or GCC enriched in the 5′ UTR, CAG trinucleotides in the CDS coding for polyglutamine tracts, and CTG triplets in the 3′ UTR. Motifs of unstable intronic microsatellites have diverse sequences [23].

Pathogenic expanded microsatellite alleles in genes are often named full mutation in comparison to shorter alleles that can comprise the classes of normal alleles or an intermediate class of premutation. Normal alleles reflect the length polymorphism present in the general unaffected population, commonly with a wide range of variation. Full mutations are considered beyond a certain upper limit of premutations, depending on the location of the microsatellite, with CDS CAG repeats showing the lowest boundaries (Fig. 6.2). Finally, the lower limits of premutations will depend on the mitotic stability of normal alleles, often defined by the microsatellite motif and its location. Common distributions of normal, premutated and fully mutated alleles can be observed in Fig. 6.2 for 5′ UTR CGG repeats, CDS CAG repeats, GAA intron repeats, and 3′ UTR CTG repeats.

The dynamic character of microsatellite instability has named the expanded unstable repeats as dynamic mutations. Normal alleles are usually stable in mitosis and meiosis. Although, in general, premutations do not associate with disease (one exception is *FMR1* premutations; see Sect. 6.4.1.1), they present meiotic instability with high probability for expansion, transmitting a full mutation to the next generation. Full mutations may produce somatic mosaicism as they are mitotically

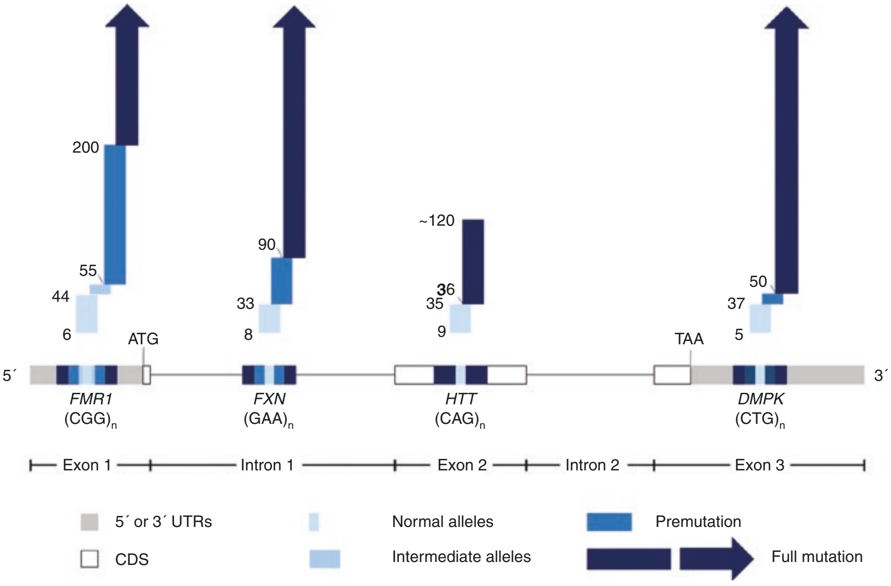


Fig. 6.2 Diagram of gene segments, microsatellite motifs and allele classes in four groups of unstable repeats associated with disease. A diagram of a hypothetical gene with three exons and two introns is presented. Translation start (ATG) and termination (TAA) codons are indicated, as well as the coding sequence (CDS), 5′ and 3′ untranslated regions (UTR). The microsatellite prototypes selected are 5′ UTR $(CGG)_n$ in the *FMR1* gene, intronic $(GAA)_n$ in *FXN*, CDS $(CAG)_n$ in *HTT* and 3′ UTR $(CTG)_n$ in *DMPK*, showing the allele classes of normal, premutation and full mutation, and respective repeat number ranges related to the pathologies: fragile X syndrome (FXS), Friedreich ataxia (FRDA), Huntington disease (HD) and type 1 myotonic dystrophy (DM1), respectively. *FMR1* has also the intermediate allele class

unstable, possibly experiencing somatic expansions or contractions. This dynamic nature of microsatellite expansion mutations has been also related to the severity of the disease. Long alleles may suffer intergeneration expansions becoming even longer. For many dynamic mutation diseases, depending on their pathophysiology mechanisms, longer alleles may associate with more severe phenotypes. These two observations combined lead to the paradigm of anticipation, which applies to certain diseases such as myotonic dystrophy (see Sect. 6.4.2), and predicts that in each generation the disease can have earlier onset and or more severe manifestation if the parental fully mutated allele has undergone further expansion.

Full mutations in distinct genes have been associated mostly with neurological disorders. In general, microsatellite expansion disorders have dominant inheritance patterns, except for certain ataxias caused by intronic repeat expansion, such as Friedreich ataxia, that show autosomal recessive inheritance (see Sect. 6.4.4). Taking into consideration that full mutations arise from length expansion of unstable premutations of transcribed microsatellite that is highly polymorphic in the general population, full mutation-associated phenotypes can be classified as follows. Rare fragile sites with or without neurodevelopmental disorders presenting intellectual disability and behavior deficits are in general the manifestation of full mutations of untranslated CGG/GCC in the 5′ UTR of genes, associated with promoter cytosine methylation. Neuromuscular disorders comprehend expanded unstable alleles of (1) CGG repeats in the 5′ UTR with non-methylated cytosines; (2) CAG repeats translated into polyglutamine tracts; (3) untranslated CTG repeats in 3′ UTR; and (4) a variety of intronic repeat motifs [23]. Disorders that fit to this classification are listed in Table 6.1. However, different reports present exceptions for this classification, such as expansions of minisatellite instead of microsatellite and location of the repeats near the promoter upstream the transcription start site, and will be highlighted at the end of the following sections. Each following Sects. (6.4.1, 6.4.2, 6.4.3 and 6.4.4) will describe a prototype disease of a microsatellite motif found in a gene specific location (5′ UTR, 3′ UTR, CDS and intron, respectively).

6.4.1 5′UTR CGG/GCC Repeats and Rare Chromosomal Fragile Sites

Cells in culture when exposed to chemical reagents that induce a perturbation of the DNA replication process may develop chromosomal fragile sites, which are gaps, constrictions or breaks on metaphase chromosomes, observed by optical microscopy. According to their frequency in the general population, fragile sites classify as common or rare. Common fragile sites can be observed in all chromosomes often induced by aphidicolin or 5-azacytidine, whereas rare fragile sites appear in up to 5% of the population, being either inherited from one of the parents or originate *de novo*.

The majority of rare fragile sites develops *in vitro* under conditions of folate deficiency or inhibition of folate metabolism inducing DNA replication stress. Differently from common fragile sites that are generally coincident with AT-rich DNA regions, many folate-sensitive fragile sites have been associated with expansions of microsatellites of CGG/GCC trinucleotide motifs. These tandem repeats locate in the 5′ UTR of genes and show length polymorphism in the general population. Some CGG/GCC expansions in the 5′ UTR of genes associate with disease, mostly neurodevelopmental disorders (Table 6.1).

FRAXA (Xq27.3) was the first rare fragile site that has been related to CGG repeat expansions, which locate in the first exon of the *FMR1* gene. FRAXA had been studied as a marker that segregated linked to a mutation that cause the most common inherited form of intellectual disability among men, fragile X syndrome (FXS; MIM 300624), denominated according to the fragile site. For decades before the cloning of the *FMR1* gene, cytogenetic analysis has been performed to diagnose FXS until the identification that FRAXA was due to expansions to more than 200 CGG repeats in the 5′ UTR of *FMR1*, the allele category named full mutation. The number of *FMR1* CGG repeats varies between 6 and 44 in the general population, whereas full mutations harbor more than 200 CGG repeats and cause FXS (Fig. 6.2). *FMR1* promoter has been characterized as a CpG island (see Chap. 4) that extends

Table 6.1 Rare fragile sites and clinical conditions associated with expansions of transcribed polymorphic microsatellites

5′ UTR CGG/GCC repeat expansion associated with chromatin condensation - Fragile sites (disease)	MIM[a]	Gene[b]
FRA10A	608866	*FRA10AC1*
FRA11A	616109	*C11ORF80*
FRA11B (Jacobsen syndrome)	147791	*CBL*
FRA12A (intellectual disability)	136630	*DIP2B*
FRA22A	138981	*CSNK1E*
FRAXA (fragile X syndrome)	300624	*FMR1*
FRAXE (FRAXE intellectual disability)	309548	*AFF2*
FRAXF	300031	*TMEM185A*
5′ UTR CGG repeat expansion with no evidence of cytosine methylation		
Fragile X-associated primary ovarian insufficiency (FXPOI)	311360	*FMR1*
Fragile X-associated tremor and ataxia syndrome (FXTAS)	300623	*FMR1*
Neuronal intranuclear inclusion disease (NIID)	603472	*NOTCH2NLC*
Oculopharyngodistal myopathy 1 (OCPDM1)	164310	*LRP12*
Oculopharyngodistal myopathy 2 (OCPDM2)	618940	*GIPC1*
Oculopharyngeal myopathy with leukoencephalopathy (OPML)	618637	*NUTM2B-AS1*
3′ UTR CTG repeat expansion		
Myotonic dystrophy, type 1 (DM1)	160900	*DMPK*
Spinocerebellar ataxia type 8 (SCA8)	608768	*ATXN8*
Translated CAG repeat expansion		
Dentatorubro pallidoluysian atrophy (DRPLA; Haw-River syndrome)	125370	*ATN1*
Huntington disease (HD)	143100	*HTT*
Huntington disease-like 2 (HDL2) [CTG]/[CAG][c]	606438	*JPH3*
Spinal and bulbar muscular atrophy (SBMA)	313200	*AR*
Spinocerebellar ataxia type 1 (SCA1)	164400	*ATXN1*
Spinocerebellar ataxia type 2 (SCA2)	183090	*ATXN2*
Spinocerebellar ataxia type 3 (SCA3; Machado-Joseph disease)	109150	*ATXN3*
Spinocerebellar ataxia type 6 (SCA6)	183086	*CACNA1A*
Spinocerebellar ataxia type 7 (SCA7)	164500	*ATXN7*
Spinocerebellar ataxia type 12 (SCA12)	604326	*PPP2R2B*
Spinocerebellar ataxia type 17 (SCA17)	607136	*TBP*
Intronic repeat expansion [repeat motif]		
Trinucleotide motif		
Friedreich Ataxia (FRDA) [GAA]	229300	*FXN*
Fuchs endothelial corneal dystrophy 3 (FECD3) [CTG]	613267	*TCF4*
Tetranucleotide motif		
Myotonic dystrophy, type 2 (DM2) [CCTG]	602688	*CNBP*
Pentanucleotide motif		
Benign adult familial myoclonic epilepsy (BAFME1)/Familial adult myoclonic epilepsy (FAME1) [TTTTA and TTTCA]	601068	*SAMD12*
BAFME2/FAME2 [TTTTA and TTTCA]	607876	*STARD7*

Table 6.1 (continued)

5′ UTR CGG/GCC repeat expansion associated with chromatin condensation - Fragile sites (disease)	MIM[a]	Gene[b]
BAFME3/FAME3 [TTTTA and TTTCA]	613608	*MARCH6*
BAFME4/FAME4 [TTTTA and TTTCA]	615127	*YEATS2*
BAFME6/FAME6 [TTTTA and TTTCA]	618087	*TNRC6A*
BAFME7/FAME7 [TTTTA and TTTCA]	61875	*RAPGEF2*
Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS) [AAGGG]	614575	*RFC1*
Spinocerebellar ataxia type 10 (SCA10) [ATTCT]	603516	*ATXN10*
Spinocerebellar ataxia type 31(SCA31) [TGGAA]	117210	*BEAN1*
Spinocerebellar ataxia type 37 (SCA37) [ATTTC]	615945	*DAB1*
Hexanucleotide motif		
Amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD) [GGGGCC]	105550	*C9ORF72*
Spinocerebellar ataxia type 36 (SCA36) [GGCCTG]	614153	*NOP56*
X-linked dystonia parkinsonism (XDP) [CCCTCT]	314250	*TAF1*

[a] MIM: Inheritance in Man (https://omim.org/)

[b] Gene names: *AFF2* (AF4/FMR2 family member 2); *AR* (Androgen receptor)*; ATN1* (Atrophin1)*; ATXN1* (Ataxin 1)*; ATXN2* (Ataxin 2)*; ATXN3* (Ataxin 3)*; ATXN7* (Ataxin 7)*; ATXN8* (Ataxin 8)*; ATXN10* (Ataxin 10)*; BEAN1* (Brain expressed associated with NEDD4 1)*; C11ORF80* (Chromosome 11 open reading frame 80)*; C9ORF72* (C9ORF72-SMCR8 complex subunit)*; CACNA1A* (Calcium voltage-gated channel subunit alpha1 A)*; CBL* (CBL proto-oncogene)*; CNBP* (CCHC-type zinc finger nucleic acid binding protein)*; CSNK1E* (Casein kinase 1 epsilon)*; DAB1* (DAB adaptor protein 1)*; DIP2B* (Disco interacting protei 2 homolog B)*; DMPK* (DM1 protein kinase)*; FMR1* (Fragile Mental Retardation 1)*; FRA10AC1* (FRA10A associated CGG repeat 1)*; FXN* (Frataxin)*; GIPC1* (GIPC PDZ domain containing family member 1)*; HTT* (Huntingtin)*; JPH3* (Junctophilin 3)*; LRP12* (LDL receptor related protein 12)*; MARCHF6* (Membrane associated ring-CH-type finger 6)*; NOP56* (NOP56 ribonucleoprotein)*; NOTCH2NLC* (NOTCH2 N-terminal-like C)*; NUTM2B-AS1* (NUTM2B antisense RNA1)*; PPP2R2B* (Protein phosphatase 2 regulatory subunit Bbeta)*; RAPGEF2* (Rap guanine nucleotide exchange factor 2)*; RFC1* (Replication factor C subunit 1)*; SAMD12* (Sterile alpha motif domain containing 12)*; STARD7* (StAR-related lipid transfer domain containing 7)*; TAF1* (TATA-box binding protein associated factor 1)*; TBP* (TATA-box binding protein)*; TCF4* (Transcription factor 4)*; TMEM185A* (Transmembrane protein 185A)*; TNRC6A* (Trinucleotide repeat containing adaptor 6A)*; YEATS2* (YEATS domain containing 2)

[c] Trinucleotide repeats in an alternative coding exon. Gene has bidirectional transcription with possibility of translation of CTG repeat or CAG repeat in the complementary strand

along its first exon, which contains the CGG microsatellite. *FMR1* CpG island cytosines are methylated in fully mutated alleles, including the cytosines of the tandem repeats, coincident with gain of repressive histone marks (H3K9 and H3K27 trimethylation), epigenetic modifications thought to be acquired in developmentally regulated processes leading to the transcription repression of the gene and the ype ([23]; see Sect. 6.4.1.1). Two other allele classes have been defined according to the instability of *FMR1* repeats: intermediate (gray-zone) alleles have 45–54 CGG repeats, and premutations, 55–200 trinucleotides (Fig. 6.2). The intermediate-allele class, though absent in the allele classification of other dynamic mutation loci, has been integrated to the categories of *FMR1* repeats because the 45–54 repeat range has transitional instability. As normal alleles, intermediate alleles and premutations are non-methylated [2]. The repeat number range that classifies FRAXA normal, intermediate-premutation and full mutation (Fig. 6.2), although not exactly the same, overlaps the allele classes of other fragile sites also due to 5′ UTR CGG expansion. Of note, *FMR1* normal alleles (including the commonest alleles with 29 and 30 CGG repeats) have in general one AGG trinucleotide intercalating between series of 9 and 10 CGG repeats.

The mothers of fully mutated FXS boys are obligate carriers of a premutation. *FMR1* premutations show meiotic instability and elevated odds to expand to full mutation in the ovaries, notably if uninterrupted by AGG units (pure tracts) and with more than 90 CGG repeats. Intermediate allele sizes overlap with the upper end of normal alleles and lower end of the premutation range. Intergenerational small contractions or expansions of CGG repeats of alleles within the normal range, intermediate range or lower-end premutation have frequently lost AGG interruptions, or have a single interruption followed by a long uninterrupted stretch of CGG repeats in its 3′ end [2].

FRAXA can be observed only in metaphase X-chromosomes of carriers of the *FMR1* full mutation, as its cytogenetic expression requires the full mutation, chromatin condensation and late DNA replication. The processes of chromatin condensation and gene transcriptional repression are common to 5′ UTR CGG expansions to more than 200 repeats. Therefore, considering the known features of folate-sensitive fragile sites, novel 5′ UTR CGG expansions can be identified. A large (>20,000 individuals) whole genome survey on promoter methylation events identified 25 loci with rare hypermethylation coincident with unstable CGG tandem repeats. Eight of them overlap rare fragile site mapping: FRA1M (*ABCD3*; ATP-binding cassette subfamily D member 3), FRA2B (*BCL2L11*; BCL2-like 11), FRA5G (*FAM193B*; Family with sequence similarity 193 member B), FRA8A (*FZD6*; Frizzled class receptor 6), FRA9A (*C9ORF72*; C9ORF72-SMCR8 complex subunit), FRA19B (*LINGO3*; Leucine-rich repeat and Ig domain containing 3), FRA20A (*RALGAPA2*; Ral GTPase activating protein catalytic subunit alpha 2) and FRA22A (*CSNK1E*; Casein kinase 1 epsilon). Of note, FRA9A is coincident with the *C9ORF72* gene that has intronic hexanucleotide repeats expanded in amyotrophic lateral sclerosis (ALS; Table 6.1). Further analysis on FRA22A-expressing individuals confirmed the CGG repeat expansion in the 5′ UTR of the gene *CSNK1E* causing this fragile site (Table 6.1). Thus, testing the other seven fragile sites for CGG repeat expansions in the respective genes is likely to add to the growing list of rare folate-sensitive fragile sites caused by unstable repeats [8].

Among folate-sensitive fragile sites previously characterized at the molecular level, expansions of CGG repeats have been also observed in gene’s first introns. FRA7A (MIM 616181) has expanded CGG repeats in the first intron of the *ZNF713* (Zinc finger protein 713) gene, related to autism spectrum disorder and transcription repression due to hypermethylation, and FRA2A (MIM 601464) corresponds to CGG expansion in an alternative CpG island promoter in the second intron of the *AFF3* (AF4/FMR2 family member 3) gene [8]. Methylation assay results have not been consistent among individuals that express FRA2A. Moreover, a condition of early onset global developmental delay with progressive ataxia and elevated glutamine (MIM 614812) has been reported in three unrelated individuals that present large expansions in the 5′UTR GCA repeats in the *GLS* (Glutaminase) gene in homozygosity or compound heterozygosity, without fragile site detection.

Few reports on repeat expansions in the 5′ UTR or upstream also differ in some aspects from the description in this section. Progressive myoclonic epilepsy type 1 (EPM1; MIM 254800; EPM of Unverricht and Lundborg type, [23]), Richieri-Costa-Pereira syndrome (RCPS; MIM 268305, [7]) and Baratela-Scott syndrome (BSS; Desbuquois dysplasia 2, MIM 615777, [14]) have autosomal recessive inheritance. EPM1 and RCPS are due to expansion of minisatellites; EPM1 and BSS relate to non-transcribed repeats located in gene promoter upstream of the transcription start sites; and RCPS and BSS are skeletal dysplasia disorders. It is possible that, according to the tissue expression pattern of these genes and period of development affected by its loss of function, haplosufficiency should explain the functional availability of the respective mRNA and protein. Moreover, transcription start sites can be altered due to expanded repeats in the 5′ end of the gene, as reported for FXS. Hence, as these diseases are rare, it is early to assume that their promoter-associated repeats are not transcribed upon expansion (see Chap. 4).

EPM1 is a rare condition, with childhood to adolescence onset that occurs more commonly in Finland and western Mediterranean. It is characterized by stimulus-sensitive myoclonus and tonic-clonic seizures that may progress to intellectual disability and cerebellar ataxia. EPM1 is most frequently due to expansion of a 12-mer minisatellite repeat in the promoter region of the *CSTB* (Cystatyn B) gene. While normal alleles contain 2–3 copies of the minisatellite motif (CCCCGCCCCGCG), affected individuals harbor alleles with 30–78 repeats in homozygosity, reducing the amount of *CSTB* mRNA. Premutations with 12–17 repeats, although non-penetrant, are unstable and prone to intergenerational expansion [23, 24].

RCPS, a craniofacial disorder associated with limb defects described in Brazilian patients, is due to expansion of 18- to 20-mer minisatellite sequence in the 5′ UTR of the *EIF4A3* (Eukaryotic translation initiation factor 4A3) gene, from 5 to 12 repeats in the normal range to 14 or 16 repeats in homozygosity in patients. Expanded alleles associate with transcription attenuation [7]. An additional example not entirely concordant with the model for microsatellite expansions is BSS, a rare disorder characterized by short stature with skeletal dysplasia, facial dysmorphisms, and developmental delay. It has been recently described that the most common genetic cause of this syndrome is CGG repeat expansions in the promoter of the *XYLT1* (Xylosyltransferase 1) gene, associated with hypermethylation of its cytosines and transcription repression [14]. This expansion is coincident with the one reported in 1995 for FRA16A fragile site [19]. These three examples illustrate that not only microsatellite may undergo expansions in the 5′ UTR of genes causing disease, but also minisatellites, and that repeats located upstream of the transcription start site can also affect transcription.

Moreover, expansion of non-transcribed minisatellite sequences has been associated with the rare fragile site FRA16B. The cytogenetic expression of FRA16B

(MIM 136580) is sensitive to chemicals that bind to AT-rich regions of DNA, and this clinically unrelated fragile site is due to the expansion of a 33-bp minisatellite repeat at 16q22.1 to 15–70 kb in size. As this expansion is present in heterozygosity in approximately 5% of individuals with European descent, FRA16B is believed to be the most common among rare fragile sites.

6.4.1.1 One Family, Three Generations, Three Distinct Clinical Conditions

Expansions of CGG trinucleotides in the 5′ UTR of the *FMR1* gene may associate with at least three clinically distinct conditions in the same family: FXS, fragile X-associated primary ovarian insufficiency (FXPOI; MIM 311360) and fragile X-associated tremor and ataxia syndrome (FXTAS; MIM 300623). While there is no clear evidence for association of intermediate alleles with a clinical phenotype, *FMR1* premutations may cause both, FXPOI and FXTAS. Moreover, a forth condition has been recently suggested to occur in individuals with premutated alleles and has received the acronym FXAND, standing for fragile X-associated neuropsychiatric disorders. Remarkably, the individuals affected by these distinct disorders do not show random positions in a FXS genealogy. FXPOI and FXTAS patients are most invariably the mother and maternal grandfather of a FXS boy, respectively. Although FXAND has not yet been widely studied, it is believed to affect different generations of premutation carriers (Fig. 6.3).

FXS is the most frequent inherited cause of intellectual disability among men. It is a clinically variable neurodevelopmental disorder affecting multiple aspects of the individual functioning. *FMR1* maps to Xq27.3 and FXS has X-linked dominant inheritance pattern. Due to X-chromosome inactivation in females (see Chap. 2), the manifestation of FXS among females tend to be relatively milder than among males. The Fragile X Online Registry With Accessible Research Database (FORWARD) integrates data from the Fragile X Clinic and Research Consortium (National Fragile X Foundation, VA, USA). Its data (78% males) refer thus only to those patients that more probably need consistent longitudinal clinical follow-up, with a trend to overrepresent more severe cases attended at a specialized clinic. In the FORWARD analysis with 362 male and 106 female individuals, carriers of the *FMR1* full mutation, older than 7 years, 99.4% and 78.3% presented intellectual disability, respectively. In comparison to female individuals, intellectual disability among males were more frequently moderate (59.9% and 12.3%), severe (12.4% and 1.9%), or profound (0.6% and 0%). Females had more borderline intellectual disability (31.1%) than males (3.6%), and mild cases ranged from 22.9% (females) to 33% (males). Among 162 male and 41 female individuals, carriers of a full mutation, under 7 years of age, development delay was present in all males and 71% females [26].

Most FXS patients meet the diagnostic criteria for attention-deficit/hyperactivity disorder (ADHD), presenting attention problems, hyperactivity and or

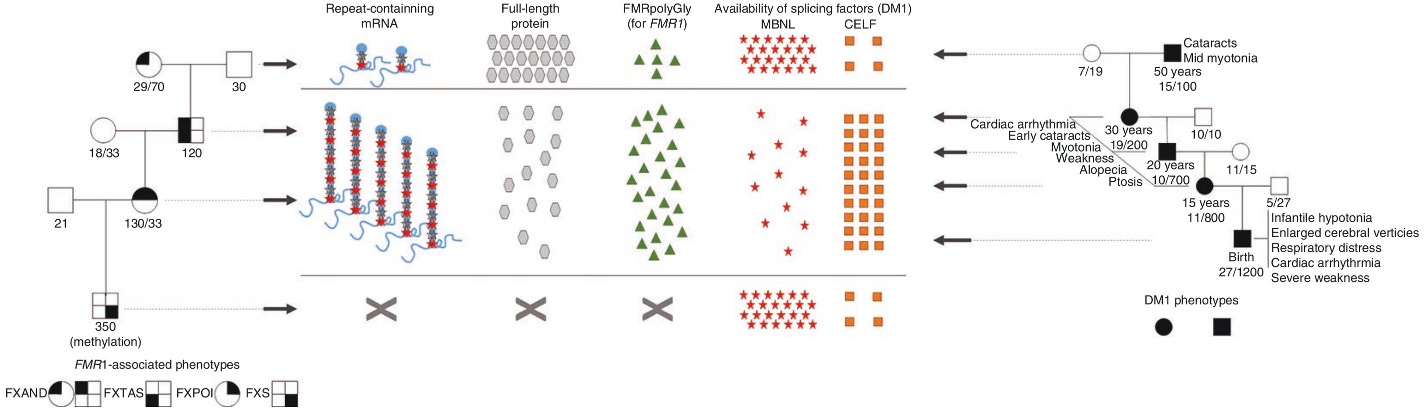


Fig. 6.3 Two illustrative genealogies are presented, for the inheritance of *FMR1* (left panel) and *DMPK* (right panel) unstable alleles. *FMR1* is involved in FXS, FXPOI, FXTAS and FXAND. *DMPK* is involved in DM1, which can manifest with clinical anticipation as indicated on the right-side genealogy by possible clinical features and putative age at onset for each generation. Number of repeats is indicated in both genealogies as possible genotypes of each individual (n/n or n for *FMR1* hemyzygous males). The central panel illustrations apply individually for *FMR1* or *DMPK* normal alleles (first row) and *FMR1* premutation and *DMPK* full mutation (central row), and *FMR1*-specific full mutation (bottom row), and show the relative amounts of mRNA and full-length protein expressed by each allele class, respectively, and FMRpolyGly peptides (for *FMR1* alleles) or availability of splicing factors (for DM1) MBNL and CELF, as indicated. The same symbol for MBNL is illustrated sequestered by long hairpins formed by *DMPK* full mutation expanded repeat-containing mRNA

impulsivity, and the criteria for anxiety disorder. The clinic-based FORWARD study diagnosed autism spectrum disorders (ASD) in 49.9% (N = 224/449) and 16.9% (N = 23/136) of respectively male and female individuals, carriers of full mutations and older than 3 years of age [26]. Although earlier studies probably overestimated the contribution of FXS to the overall incidence of ASD among boys (nearly 6%), more recent studies disclosed an approximate rate of 0.5%, most likely due to better clinical recognition of ASD and genetic diagnosis employing next-generation sequencing and comparative genomic hybridizations techniques (see Chaps. 3 and 9). The comorbidity of FXS and ASD increases the severity of cognitive and behavioral problems when compared to FXS without ASD [13]. Moreover, few female (6%) and male (14%) FXS patients develop seizures that manifest as developmental-behavioral comorbidity, generally not refractory to medication [3].

Finally, general physical features are commonly present in FXS male and female individuals, such as high-arched palate, long narrow face, macrocephaly, prominent jaw, prominent ears; macroorchidism in pubertal and post-pubertal males; joint hypermobility, *pectus excavatum*, flat feet, scoliosis; aortic root dilation, mitral valve prolapse; recurrent otitis media; strabismus and refractive visual errors [26].

Primary ovarian insufficiency (POI) is defined as 6 months of amenorrhea in women under 40 years of age, presenting elevated follicle-stimulating hormone (FSH) and low estradiol levels, arising as consequence of genetic or non-genetic factors that result in decreased initial primordial follicle number, increase in follicle apoptosis or a failure of the follicle to respond to gonadotrophin stimulation. It has been estimated that nearly 16% of women with the *FMR1* premutation develop FXPOI. Notably, women without FXPOI, carriers of the *FMR1* premutation, tend to present menopause earlier than their sisters homozygous for normal alleles. In addition, a subset of women with age lower than 40 and the *FMR1* premutation may have high FSH levels and regular menses, characterizing subclinical ovarian failure [2, 29].

FXTAS is a late-onset, neurodegenerative condition that affects mostly men, characterized by progressive intention tremor, gait ataxia, cognitive decline, peripheral neuropathy and autonomic dysfunction. The mean age at onset of tremor has been reported to be nearly 62 years (39–78) and of gait ataxia 63 years of age (47–78). Cognitive decline and behavioral problems have been observed in *FMR1* premutation carriers with or without FXTAS [9, 23, 24]. *FMR1* CGG repeat length correlates with age at onset of FXTAS motor signs and cognitive decline, but not with severity, consisting an example of clinical anticipation. Evidence indicates that a subset of male and female premutation carriers may present isolated neuropsychiatric conditions (anxiety, depression, ADHD, social deficits, ASD, obsessive-compulsive disorders) that led to the proposal of FXAND, which typically has earlier onset than FXTAS, affecting both children and adults.

6.4.1.2 Genetic Pathophysiology of *FMR1* Premutation

The *postmortem* assessment of FXTAS patient central nervous system disclosed demyelination in white matter from cerebrum and cerebellum, spongiform intercellular edema in the middle cerebellar peduncles, as well as intranuclear, eosinophilic, ubiquitin-positive inclusions in neurons and astrocytes, in spinal cord, cerebellum and brain. These neuronal inclusions contain *FMR1* mRNA, and the inclusion amount correlates with *FMR1* CGG repeat number [9]. Although no significant alteration in FMRP protein levels is generally observed in FXTAS samples, but a mild decrease in its amount, *FMR1* mRNA levels are consistently increased (Fig. 6.3). This led to the proposition of a toxic RNA model, similar to the one described for myotonic dystrophy (see Sect. 6.4.2), in which CGG repeats would hamper RNA polymerase II progression, signaling to a feedback response to increase *FMR1* transcription. The RNA with the expanded repeats would bind non-specifically to regulatory proteins, originating the inclusions. The misregulation of those RNA-binding regulators would in turn disrupt cellular mechanisms, including splicing. *Fmr1* knock-in mice expressing 98 CGG repeats and *Drosophila melanogaster* mutants with long CGG repetitive tracts show high levels of *Fmr1* mRNA and neuronal inclusions, corroborating the proposed model [23].

The toxic gain-of-function RNA model is further supported by a recognized unconventional mode of translation, termed repeat-associated non-AUG-initiated (RAN) translation, up-regulated after *FMR1* mRNA accumulation. RAN translation applied to microsatellites involves production of poly-amino acid-containing peptides, notably for trinucleotide motifs and poly-dipeptide for hexanucleotide motifs. As the inclusions described in FXTAS tissue resemble inclusions from protein gain-of-function-mediated neurodegenerative disorders (see Sect. 6.4.3), *FMR1* RAN translation has been demonstrated starting upstream of the CGG repeats, expressing a peptide with predicted mass of 11.5 kDa. The peptide, named FMRpolyGly, contains an N-terminal poly-glycine (polyGly) stretch followed by a 42-amino acid carboxyl terminal domain, out of frame with the downstream FMRP start codon. That is thus considered an upstream open reading frame (uORF) that is actually translated (see Chap. 4). Although FMRpolyGly can be expressed in low quantities by *FMR1* common-intermediate alleles, its amount is significantly increased in cells with a premutation. In normal-intermediate alleles, CGG uORF should serve as a translational control of the main ORF. FMRpolyGly has been described in FXTAS intranuclear inclusions, and appears to induce ubiquitin-proteasome system impairment in insect cells [20, 24].

Studies indicated a non-linear relationship between FXPOI risk and CGG repeat number within the premutation range, with 89 CGG repeats correlating more strongly with FXPOI risk than lower or higher ends of the premutation class. In addition, altered menstrual cycle traits, subfertility and dizygotic twinning associate with alleles with 80–99 repeats [2]. A knock-in mouse model expressing *Fmr1*

$(CGG)_{90}$ premutation mRNA showed reduced number of growing follicles in ovaries [17], while the expression of an allele with 130 repeats by another mouse model associated with faster loss of follicles [10]. These mice have a normal primordial follicle pool, which is depleted more quickly than in the wild-type mice. Therefore, two molecular models have been proposed: one mediated by $(CGG)_{80–99}$ causing reduction of fetal ovarian follicle numbers and another with $(CGG)_{100–200}$ associated with follicle atresia [25]. The observation of nuclear ubiquitin- and FMRpolyGly-positive inclusions in the ovary of a FXPOI patient indicates that FXPOI and FXTAS should arise by similar pathogenic mechanisms [6].

Recent works have based on the knowledge of the pathophysiology of FXTAS and its clinical and pathological similarities with neuronal intranuclear inclusion disease (NIID) to search for CGG repeats causing this pathology. CGG expansions in the 5′ UTR of the *NOTCH2NLC* (Notch2 N-terminal-like C) gene to the size observed in *FMR1* premutations (90–180) were detected in NIID patients, whereas control individuals had 9–43 repeats (Table 6.1) [12].

6.4.2 3′UTR CTG repeats and Type 1 Myotonic Dystrophy

Type 1 myotonic dystrophy (DM1) is the most common form of adult muscular dystrophy and is caused by expansion of CTG repeats in the 3′ UTR of the *DMPK* (DM1 protein kinase) gene. Type 2 myotonic dystrophy (DM2), less frequent and severe than DM1, also results from unstable microsatellite expansion but of tetranucleotide CCTG repeats in intron 1 of the *CNBP* (CCHC-type zinc finger nucleic acid binding protein) gene (Table 6.1). DM1 and DM2 comprise a group of multisystem diseases with autosomal dominant inheritance, characterized by the core features of myotonia, muscle weakness, muscular dystrophy, early-onset cataracts (before 50 years), cardiac conduction defects and endocrine disorders [27].

Genetic anticipation is classically observed in DM1 (Fig. 6.3). The clinical findings between ages 20 and 70 years for carriers of alleles with 50–100 repeats generally include cataracts before age 50 and mild myotonia. Individuals with 10–30 years and 50–1,000 repeats may present early cataracts, distal weakness, myotonia, temporomandibular wasting, ptosis, balding, excessive daytime sleepiness, central obstructive sleep apnea, respiratory failure, cardiac conduction abnormalities, insulin resistance, mood disorders, and or mild intellectual deficits. DM1 with childhood onset (1–10 years; 50–1,000 repeats) may aggravate myotonia with facial weakness and dysarthria, cardiac conduction defects as well as intellectual disability. Congenital DM1 (>1,000 repeats) may present prenatally with polyhydramnios, and at birth with hypotonia, severe weakness, respiratory failure, cardiopulmonary complications, cerebral atrophy, and enlarged cerebral ventricles. In addition, a bias towards maternal transmission of expanded alleles has been observed in DM1, a feature common to other unstable repeat disorders [27].

The molecular pathophysiology of DM1 is similar to the one described for FXTAS mediated by intracellular accumulation of mRNA produced by the gene that harbors the trinucleotide repeats in nuclear ubiquitin-positive inclusions and a toxic gain of function [15]. In fact, this model was first established for DM1-associated *DMPK* expanded CTG repeats and later tested in other diseases that course with increased repeat-containing mRNA and intracellular inclusions (see Sect. 6.4.1.2) [12]. Therefore, the mechanistic model of DM1 is the prototype of RNA gain-of-function pathogenesis, and has been considered and tested in spinocerebellar ataxia (SCA) 8 (SCA8), which is also due to repeat expansion in the 3′ UTR of a gene (*ATXN8*), and other pathologies of untranslated repeats as intron repeat expansion diseases (SCA31, SCA36, FECD3, ALS/FTD, and DM2), and HDL2, besides FXTAS and NIID (Table 6.1; Sect. 6.4.1.2). It is relevant to highlight that HDL2 is owing to CTG repeats in an alternative exon of *JPH3*. When skipped the exon is considered an intron, as the length extension may weaken its recognition as exon by the spliceosome. Thus, if HDL2 is mediated by mRNA toxicity it could possibly be by an intron-related mechanism. Although it is well accepted that DM1 major mechanism is mediated by *DMPK* mRNA (see below), relative reduction of the encoded protein has been in general observed for *DMPK* and several other unstably mutated genes. Moreover, while proteotoxicity and RAN translation are not believed to play major roles in DM1, they are widely accepted for the pathogenesis of diseases such as ALS/FTD, in addition to the RNA toxicity [23, 24].

Nuclear proteins playing regulatory roles in pre-mRNA splicing have been shown to be entrapped in *DMPK* mRNA-containing inclusions. This model has been extensively tested in DM1 tissues and experimental systems, and identified Muscleblind (MBNL) family members as major proteins sequestered into nuclear inclusions, and upregulation of CUG RNA-binding protein Elav-like family member 1 (CELF1), two important splicing regulators (Fig. 6.3). MBNL activity appears key in maintaining the expression of mRNA encoding proteins with essential roles for the adult muscle phenotype. The combinatorial effect of down-regulated MBNL and up-regulated CELF results in a muscle expression profile consistent with the fetal stage leading to loss of cell viability and function [23, 24].

Different therapeutic methods have been tested in animal models for the major uncontrolled targets related to DM1: RNA interference (RNAi) to reduce *DMPK* mRNA levels; chemicals as oligonucleotide or morpholino that block the CUG-repeat RNA or chimeric site-specific oligonucleotides that recruit RNAse H cleaving the repeat sequence; antisense oligonucleotide for splice site selection regulation in MBNL/CELF-specific pre-mRNA target relevant to control myotonia; and restoration of MBNL function by miRNA sponge or antagomir interfering in its related microRNA [21]. It is expected that successful therapeutic preclinical trials will move ahead into clinical trials. Oligonucleotide-based therapies recently approved for spinal muscular atrophy (see Chap. 4) and Duchenne muscular dystrophy are highly inspiring and motivating for the research field of DM1 and other unstable repeat disorders mediated by RNA toxicity.

6.4.3 Translated CAG Repeats

Common features of unstable translated CAG alleles are the dominant inheritance pattern and the synthesis of the encoded protein expressing extended polyglutamine tracts prone to form insoluble aggregates beyond 35–40 repeats. Aggregated polyglutamine proteins are seen as cytoplasmic or nuclear inclusions that are positive for ubiquitin suggestive of unfolded proteins. Impairment in ubiquitin-proteasome pathway as well as mitochondria dysfunction are also frequent features. In different study systems, expression of large polyglutamine expansions without the whole protein context is sufficient to elicit inclusion formation and neurodegeneration [23]. Moreover, RAN translation was demonstrated for the *HTT* (Huntingtin) gene, mutated in Huntington disease (Table 6.1), producing peptides in different frames by the sense (CAG) or antisense (CTG) strands. Preliminary data show that polyglutamine, polyserine, polyalanine, polycysteine or polyleucine peptides enhance early in HD brain inclusions and disrupt nucleocytoplasmic transport. Polyglutamine inclusions in addition to the encoded protein-specific dysfunctions associate in the pathogenesis of unstable translated CAG repeats [24].

An additional mechanism related to the expression of CDS CAG repeats has been described for the *CACNA1A* (Calcium voltage-gated channel subunit alpha1 A) gene, which contains CDS CAG repeats, expanded in SCA6 patients. Its mRNA is bicistronic, encoding the pore-forming calcium channel expressed in cerebellum through canonical, cap-dependent translation, and a novel transcription factor containing polyglutamine (pQ-TF) by internal ribosome entry site (IRES)-mediated translation from at least one alternatively spliced transcript (see Chap. 4). In search for a therapeutic strategy for SCA6, silencing the full-length *CACNA1A* mRNA would result in the loss of the calcium channel in the cerebellum. Hence, the IRES controlling the pQ-TF expression arises as a plausible candidate target for selective silencing as a therapeutic intervention for SCA6 [22].

A distinct subset of genes has trinucleotide repeats coding for polyalanine tracts of nuclear proteins, mostly transcription factors. Expansion of polyalanines in the protein associates with disease. However, differently from the loci related to the dynamic mutation clinical phenotypes mentioned so far, the alanine codons are not polymorphic in unaffected individuals. The expanded polyalanine tracts should result most likely from a short DNA segmental duplication in heterozygosity extending the region of codons for alanine. Nevertheless, the pathophysiological mechanisms leading to disease in polylanine expansion conditions appear similar to those of expanded polyglutamine-mediated disorders [11]. Some known polyalanine expansion disorders and their respective mutated genes (in brackets) include: synpolydactyly type II (MIM 186000; *HOXD13* - Homeobox D13 - gene); hand-foot-genital syndrome (MIM 140000; *HOXA13* - Homeobox A13 - gene); early infantile epileptic encephalopathy with suppression burst (Ohtahara syndrome, MIM 308350) or infantile epileptic-dyskinetic encephalopathy without obvious brain malformation (Partington syndrome, MIM 309510) both involving the *ARX* - Aristaless-related homeobox - gene; congenital central hypoventilation syndrome (MIM

209880; *PHOX2B* - Paired-like homeobox 2B - gene); X-linked intellectual disability with hypopituitarism (MIM 300123; *SOX3* - SRY-box transcription factor 3 - gene); holoprosencephaly (MIM 186000; *ZIC2* - Zic family member 3 - gene); cleidocranial dysplasia (MIM 119600; *RUNX2* - RUNX family transcription factor 2 - gene); blepharophimosis– ptosis–epicanthus inversus syndrome (MIM 110100; *FOXL2* - Forkhead box 2 - gene); and oculopharyngeal muscular dystrophy (MIM 164300; *PABPN1* - Poly-A binding protein nuclear 1 - gene).

6.4.4 Expansion of Intronic Tri-, Tetra-, Penta- and Hexanucleotide Repeats

Recently, intronic microsatellite expansions have been frequently reported for the etiology of diseases since its first association with a neurologic disease (Friedreich ataxia, FRDA) in 1996, constituting today the gene location with more known examples of disease-associated unstable repeats, including six forms of familial myoclonic epilepsy, three types of SCAs, and the most common genetic cause of ALS. There are different microsatellite motifs with tri-, tetra-, penta- and hexanucleotides involved in unstable intronic expansions (Table 6.1). This diversity of motifs and clinical associations is probably due to lesser evolutionary constraints in introns, their larger sizes (25% of the human genome, see Chap. 1), high rates of tandem repeat homing (see Sect. 6.3), and regulatory functions in splicing (see Chap. 4). Remarkably, in eight out of 16 intronic loci listed on Table 6.1 the microsatellite is within the gene's first (N = 7) or second (N = 1) introns, which should possibly have additional transcription regulatory roles. It is also important to emphasize that in the *TAF1* (TATA-box binding protein associated factor 1) gene the hexanucleotides are part of a retroelement SVA (see Sect. 6.3 and Chap. 8) in intron 32, causing X-linked dystonia Parkinsonism, a form of dystonia observed with high frequency and founder effect in the Panay population, Philippines [4]. Association with transposable elements is a common feature of tandem repeats, also reported for other unstable repeat loci described here such as those involved in myoclonic epilepsy. Furthermore, fragile sites FRA2A and FRA7A have unstable repeats in gene's second and first introns, respectively, but have not been listed on Table 6.1, because they do not follow the classification criteria adopted here (see Sect. 6.4.1). Nevertheless, it should be noticed that for the vast majority of the diseases recognized today as the product of unstable repeat expansions (nearly 70%) the microsatellite (or minisatellite; see Sect. 6.4.1) resides in the 5′ part of the gene, from the promoter associated with the gene (mostly CpG islands) to the first or second introns. This generalization logically includes repeats upstream of the promoter that are not necessarily transcribed (see Sect. 6.4.1). However, bidirectional transcription (see Chaps. 4 and 5) has been a recurrent theme in the field of dynamic mutations generating coding or noncoding transcripts (e.g., ALS/FTD, HDL2, OPML, SCA8; see Table 6.1). Yet, there is not enough evidence to suggest if repeat-containing anti-sense transcripts are beneficial or aggravate the pathology. As the first intron of the gene is generally the largest and considerably long, tandem repeats in the 5′ extended portion of the gene can be involved in several steps of gene expression, as transcription, chromatin and splicing regulation, possibly also recursive splicing of the first intron.

Among the diseases of intronic microsatellie expansions, FRDA and cerebellar ataxia, neuropathy, and vestibular areflexia (CANVAS) are two ataxia disorders with autosomal recessive inheritance, that manifest owing to expansions of trinucleotide GAA (*FXN* gene) and pentanucleotide AAGGG (*RFC1* gene), respectively. Clinical anticipation is observed in FRDA genealogies, as patients with the classical form present before age 25 neurological signs of progressive ataxia, sensory loss and areflexia, and individuals with short GAA expansions should not manifest the disease until later when spasticity and hyperflexia, instead of areflexia, are more common. Additionally, *FXN* GAA repeat length shows a direct correlation with the probability of significant cardiac dysfunction [23].

Similar to FXS, FRDA manifests because of loss of function of its protein product. In the FRDA case, there is partial loss of frataxin, which is required for mitochondria assembly of iron-sulphur clusters, preventing iron accumulation. Consequently, cells with *FXN* alleles with expanded GAA repeats in homozygosity show mitochondria dysfunction, decreased ATP production and increased levels of reactive oxygen species in the nervous system and heart. Long tracts of GAA repeats directly interfere in *FXN* transcription initiation and elongation or induce epigenetic silencing. Importantly, stable DNA-RNA hybrids containing the GAA repeats have been reported *in vitro* during *FXN* transcription elongation, significantly reducing the amount of frataxin [24]. Decreasing oxidative stress and iron accumulation, and improving mitochondria function are pharmacological targets for FRDA patients. Moreover, experimental genetic-based approaches aim at increasing cellular frataxin.

6.5 Final Remarks

Satellite DNA and tandem DNA repeats are frequently misassembled or missing in the human reference genome, and overlooked in disease association studies. Twenty years after the publication of the first human genome sequence draft, finishing it by bioinformatics and sequencing based on long-read alignment tools is now a reality to obtain complete, telomere-to-telomere sequences of human autosomes [18].

Similar strategies with limited samples of human individuals have recently clarified different loci of human genome tandem repeats, by obtaining nearly 17,500 haplotype-resolved STRs or VNTRs with respect to their orthologous alleles in three non-human primates. Nearly 1,400 non-redundant loci have been identified as human-specific expansions of tandem repeats (N = 1,021) and/or tandem repeats without evidence for their presence in non-human primates (N = 436). These loci were commonly enriched in C and G, intronic regions, and VNTRs (70%). The tandem repeats are near genes with biological function overrepresented in the central nervous system. Moreover, based on the premise that longer and purer tracts of microsatellites are more likely to undergo repeat expansions and potentially cause disease, the identification of such alleles by long-read sequencing may identify candidate loci for genetic instability [28].

The last decade has also brought considerable progress in understanding the pathophysiological mechanisms of neurological disorders caused by microsatellite expansions, remarkably for those mediated by RNA toxic gain of function and the description of RAN translation. Clinical and pathological similarities among neurological diseases have recently allowed identifying the genetic etiology of other conditions by DNA long-read sequencing and analysis centered on specific repeat motifs. It is expected that disease etiology discovery will accelerate in coming years as various novel candidate loci have been recently disclosed. Finally, the pathogenetic models established are the foundation for pharmacological and molecular biology-based interference strategies in search for novel treatments.

References

1. Miga KH, Koren S, Rhie A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585:79–88.
2. Allen EG, Sullivan AK, Marcus M, et al. Examination of reproductive aging milestones among women who carry the FMR1 premutation. Hum Reprod. 2007;22(8):2142–52.
3. Berry-Kravis E, Raspa M, Loggin-Hester L, et al. Seizures in fragile X syndrome: characteristics and comorbid diagnoses. Am J Intellect Dev Disabil. 2010;115(6):461–72.
4. Bragg DC, Sharma N, Ozelius LJ. X-linked dystonia-parkinsonism: recente advances. Curr Op Neurol. 2019;32(4):604–9.
5. Britten RJ, Kohne DE. Repeated sequences in DNA. Science. 1968;161(3841):529–40.
6. Buijsen RAM, Visser JA, Kramer P, Severijnen EAWFM, et al. Presence of inclusions positive for polyglycine containing protein, FMRpolyG, indicates that repeat-associated non-AUG translation plays a role in fragile X-associated primary ovarian insufficiency. Hum Reprod. 2016;31(1):158–68.
7. Favaro FP, Alvizi L, Zechi-Cleide RM, et al. A noncoding expansion in *EIF4A3* causes Richieri-Costa-Pereira syndrome, a craniofacial disorder associated with limb defects. Am J Hum Genet. 2014;94:120–8.
8. Garg P, Jadhav B, Rodriguez OL, et al. A survey of rare epigenetic variation in 23,116 human genomes identifies disease-relevant epivariations and novel CGG expansions. Am J Hum Genet. 2020;107:654–69.
9. Greco CM, Berman RF, Martin RM, et al. Neuropathology of fragile X-associated tremor/ataxia syndrome (FXTAS). Brain J Neurol. 2006;129(Pt. 1):243–55.
10. Hoffman GE, Le WW, Entezam A, et al. Ovarian abnormalities in a mouse model of fragile X primary ovarian insufficiency. J Histochem Cytochem. 2012;60(6):439–56.
11. Hughes JN, Thomas PQ. Molecular pathology of polyalanine expansion disorders: new perspectives from mouse models. In: Hatters D, Hannan A, editors. Tandem repeats in genes, proteins, and disease. Methods in molecular biology (methods and protocols), vol. 1017. Totowa, NJ: Humana Press; 2013.
12. Ishiura H, Tsuji S. Advances in repeat expansion disease and a new concept of repeat motif-phenotype correlation. Curr Opin Genet Dev. 2020;65:176–85.

13. Kaufmann WE, Kidd SA, Andrews HF, et al. Autism spectrum disorder in fragile X syndrome: Cooccurring conditions and current treatment. Pediatrics. 2017;139(Suppl. 3):S194–206.
14. LaCroix AJ, Stabley D, Sahraoui R, et al. GGC repeat expansion and exon 1 methylation of *XYLT1* is a common pathogenic variant in Baratela-Scott syndrome. Am J Hum Genet. 2019;104:35–44.
15. Lee JE, Cooper TA. Pathogenic mechanisms of myotonic dystrophy. Biochem Soc Trans. 2009;37(Pt. 6):1281–6.
16. Lizardo DY, Kuang C, Hao S, et al. Immunotherapy efficacy on mismatch repair-deficient colorectal cancer: from bench to bedside. BBA Rev Cancer. 2020;1874:188447.
17. Lu C, Lin L, Tan H. Fragile X premutation RNA is sufficient to cause primary ovarian insufficiency in mice. Hum Mol Genet. 2012;21(23):5039–47.
18. Miga KH. Centromere studies in the era of 'telomere-to-telomere' genomics. Exp Cell Res. 2020;394(2):112127.
19. Nancarrow JK, Holman K, Mangelsdorf M, et al. Molecular basis of p(CCG)n repeat instability at the FRA16A fragile site locus. Hum Mol Genet. 1995;4:367–72.
20. Oh SY, He F, Krans A. RAN translation atCGGrepeats induces ubiquitin proteasome systemimpairment in models of fragile X-associated tremor ataxia syndrome. Hum Mol Genet. 2015;24(15):4317–26.
21. Overby SJ, Cerro-Herreros E, Llamusi B, et al. RNA-mediated therapies in myotonic dystrophy. Drug Discov Today. 2018;23:2013–22.
22. Pastor PDH, Du X, Fazal S, et al. Targeting the *CACNA1A* IRES as a treatment for spinocerebellar ataxia type 6. Cerebellum. 2018;17:72–7.
23. Paulson H. Repeat expansion diseases. Handb Clin Neurol. 2018;147:105–23.
24. Rodriguez CM, Todd PK. New pathologic mechanisms in nucleotide repeat expansion disorders. Neurobiol Dis. 2019;130:104515.
25. Sherman SL, Curnow EC, Easley CA, et al. Use of model systems to understand the etiology of fragile X-associated primary ovarian insufficiency (FXPOI). J Neurodev Disord. 2014;6(1):26.
26. Sherman SL, Kidd SA, Riley C, Berry-Kravis E, et al. FORWARD: a registry and longitudinal clinical database to study fragile X syndrome. Pediatrics. 2017;139(Suppl. 3):S183–93.
27. Smith CA, Gutmann L. Myotonic dystrophy type 1 management and therapeutics. Curr Treat Opt Neurol. 2016;18(12):52.
28. Sulovari A, Li R, Audano PA, et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc Natl Acad Sci U S A. 2019;116(46):23243–53.
29. Vianna-Morgante AM, Costa SS, Pares AS, et al. FRAXA permutation associated with premature ovarian failure. Am J Med Genet. 1996;64(2):373–5.

Chapter 7
Human Chromosome Telomeres

Florencia Barbé-Tuana, Lucas Kich Grun, Vinícius Pierdoná, Beatriz Cristina Dias de Oliveira, Stephany Cacete Paiva, Mark Ewusi Shiburah, Vítor Luiz da Silva, Edna Gicela Ortiz Morea, Verônica Silva Fontes, and Maria Isabel Nogueira Cano

7.1 Introduction

The discovery that telomeres, the terminal structures of eukaryotic chromosomes, have a role in controlling genomic stability and reflect an individual's life experiences adds an appeal for the study of these genomic sequences. Telomeres can work as mitotic clocks and sensors of individuals' general health.

Telomeres are ribonucleoprotein complexes ranging in size from 10,000 to 15,000 base pairs (bp) located at the end of linear chromosomes. They are formed by double-stranded non-coding repeated DNA sequences arranged in tandem,

F. Barbé-Tuana (✉)
Postgraduate Program in Cellular and Molecular Biology, School of Health, Sciences and Life, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Rio Grande do Sul, Brazil

Laboratory of Immunobiology, School of Sciences, Life and Health, Pontifical Catholic University of Rio Grande do Sul – PUCRS, Porto Alegre, Brazil
e-mail: florencia.tuana@pucrs.br

L. K. Grun
Postgraduate Program in Pediatrics and Child Health, School of Medicine, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Rio Grande do Sul, Brazil
e-mail: lucas.grun@pucrs.br

V. Pierdoná
Postgraduate Program: Biochemistry, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil

B. C. D. de Oliveira · S. C. Paiva · M. E. Shiburah · V. L. da Silva · E. G. O. Morea · V. S. Fontes · M. I. N. Cano
Department of Chemical and Biological Sciences, Biosciences Institute, Sao Paulo State University, Botucatu, Brazil
e-mail: bcd.oliveira@unesp.br; edna.gicela-ortiz-morea@unesp.br; vs.fontes@unesp.br; maria.in.cano@unesp.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_7

which in mammals is composed of the TTAGGG hexamer. At the final 3′ portion, telomeres are composed of a guanine-rich single-stranded protrusion of 150–200 nucleotides called the 3′G-overhang that forms a three-dimensional structure (t-loop) maintained by the shelterin protein complex. Shelterin has regulatory functions over the telomeric length, repressing the pathways that induce a local DNA damage response (DDR) and DNA recombination. In this way, telomeres maintain cell integrity and stability, performing crucial genome protective functions.

Telomeres shorten in each cell division cycle due to the incomplete replication of linear DNA molecules, a phenomenon called "the end-replication problem." The telomerase enzyme compensates for the loss of telomeric DNA by adding new TTAGGG sequences to every short telomere, in cell populations where the enzyme is expressed. However, in most adult somatic cells, telomerase activity is suppressed. The physiological shortening of the telomeres occurs within each cell division cycle, reflecting the proliferative history of the cell. The progressive shortening of telomeres can impair the tissue's regenerative capacity and has been proposed as a representative feature of aging. Critically short telomeres induce DNA repair signaling pathways and, if not repaired, can cause cell death by apoptosis or senescence.

In this chapter, we will cover the sequence and structure of human telomeres. We will discuss its unique synthesis and regulation from a molecular point of view. We will reveal how its length, not its sequence, reflects the biological rhythm at the cellular level, and how its shortening is related to the appearance of diseases present in aging. In this context we will describe the telomeropathies, rare human syndromes caused by mutations in telomeric proteins and in components of the telomerase ribonucleoprotein structure, and how these mutations induce telomeric dysfunction and cause chronic diseases. Finally, this chapter describes therapeutic approaches targeting telomerase inhibition in tumoral cells.

7.2 History

Early observations marked the decade of the 1930s for the field of what we now know as telomeres biology. Two independent scientists, Barbara McClintock and Herman Muller, came simultaneously to the same conclusion; ionizing irradiation induces brakes along the chromosomes, but not at the ends. Occasionally, they observed that chromosome ends fused to each other or to broken chromosomes, making them special chromosomic structures, called by Muller telomeres (from the Greek: telos—terminus and meros—part). According to Muller, telomeres contained special heterochromatin that helps to cap chromosome ends [1, 2]. Another interesting result came decades later when Elizabeth Blackburn and Joseph Gall described, initially in the ciliated protozoa *Tetrahymena thermophila*, that the DNA sequence forming the ends of the chromosomes was composed of tandem repeats of 5′-TTGGGG-3′ [3]. Also intriguing was that these short repeated sequences were conserved present at the chromosomes ends in different species, independent on the upstream sequence. From the discovery of telomeres, several studies emerged intending to investigate the function of this structure. In 1982, Jack Szostak confirmed the conservation of telomere function throughout the evolution of the species [4].

In the 1960s, Leonard Hayflick observed that cultured human fibroblasts had limited capacity for division and proliferation, defined as the "Hayflick limit", which expresses the number of times normal cells can divide, directly related to senescence [5]. Few years later, studies on the properties of the DNA semiconservative replication process [6, 7] revealed particularities that corroborated Hayflick's theory of senescence. Though, the observations that progenitor cells had longer telomeres than somatic cells [8], telomeres shortened with age [9] and that tumoral cells had short telomeres [10], suggested that an enzyme capable of elongating telomeres would exist. In 1988 Carol Greider mentored by Elizabeth Blackburn discovered an enzyme with terminal transferase activity in extracts from *T. thermophila* [11] and together with Jack Szostack were honored with the Nobel Prize in Physiology and Medicine (2009) for their essential contributions to leverage the telomeres biology field (https://www.nobelprize.org/prizes/medicine/2009/summary/). Cloning of the telomerase gene and other studies on its complex structure composed by the catalytic protein component TERT [12] and RNA template (denominated TR, hTR or TERC) revealed they are conserved in structure and function among a variety of different organisms. Mutations in the human telomerase holoenzyme (hTERT or hTRT) are related to monogenic inherited diseases called primary telomeropathies [13]. The discovery that telomeropathies share defects in telomeres maintenance machinery and induce short telomeres compared to age-matched subjects opened a new field in genetics. Primary telomeropathies predispose individuals to a premature aging phenotype, with early onset and genetic antecipating. Of note, the majority of cancer cells reactivate TERT to compensate for telomeric loss consequent of vigorous proliferation [14]. In the last two decades, these observations harbour enormous enthusiasm for targeting genetic, geriatric and cancer diseases.

7.3 Human Telomeres: Maintenance for Genome Stability and Cell Proliferation

7.3.1 Telomeres and Their Structural Features

In most eukaryotes, with few exceptions, the chromosome's end has a particular tridimensional arrangement called telomere that plays essential cellular functions. They prevent chromosomes from being recognized as DNA double-strand breaks (DSB), protecting them from recombination, and terminal fusions. In this way, telomeres prevent genome instability, cell cycle arrest, and senescence [15].

Telomeres are composed of repetitive DNA in double and single-strand forms. In humans and other vertebrates, telomeres are formed of tandemly repeated six nucleotides sequence (5′-TTAGGG-3′)$_n$ about 3–15 kb in length (Fig. 7.1). They are associated with specific telomeric-interacting proteins called the shelterin complex, forming a dynamic nucleoprotein structure [16–19]. The telomeric G-rich strand forms a single-strand protrusion towards the end of the chromosome known as the 3′-G overhang (Fig. 7.1a), which in humans is about 100–200 nucleotides long [16, 20]. This single-stranded region serves as a substrate for telomere replication by the enzyme telomerase (see Sect. 7.3.2.2 for details).

Telomeres' ends can also form complex structures such as t-loop and G-quadruplex (Fig. 7.1b). T-loops are formed when telomeres' end folds back on itself. The protruded 3′G-rich single-stranded DNA is sequestered within a displacement loop into the upstream double-stranded region (the so-called D-loop), forming a lariat-like structure that shields chromosomes ends from DNA damage response (DDR) [21, 22]. The formation of t-loops is mainly stabilized by the shelterin complex, specifically the TRF2 (TTAGGG Repeat Factor 2) protein. This specialized structure hinders the 3′-hydroxyl (3′-OH) overhang from being recognized as single-strand break (SSB). It prevents the activation of the DDR machinery induced by Ataxia Telangiectasia Mutated (ATM) and Ataxia Telangiectasia Rad3-related (ATR) kinase signaling cascades [23–26]. These pathways have an

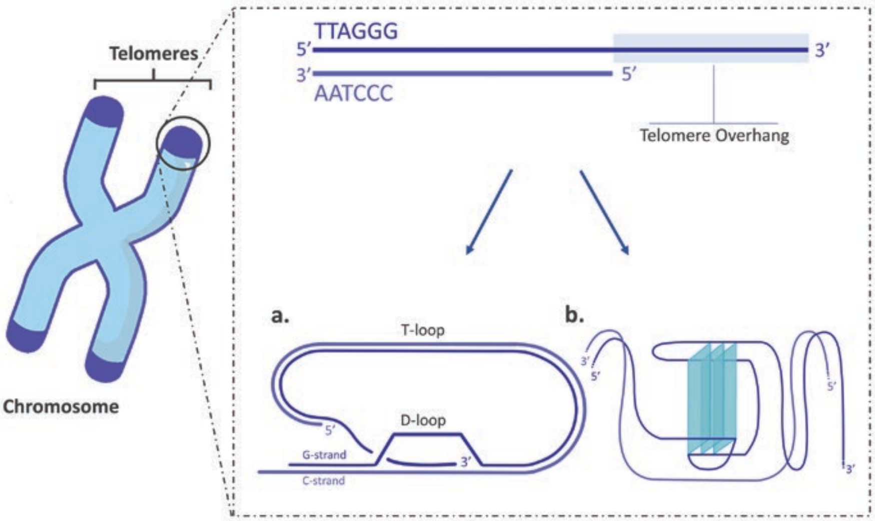


Fig. 7.1 Human telomeres can form complex structures to protect chromosome ends. Human telomeres are composed of a six-nucleotide sequence (5′-TTAGGG-3′) tandemly repeated (3–15 kb) where the G-rich strand comprises a protrusion towards the end of the chromosome known as the 3′-G overhang. (**a**) The telomeric 3′G-overhangs can form t-loops, which are lariat-like structures formed when a telomere end folds back on itself and the 3′G-overhang is sequestered within a displacement loop into the upstream double-stranded region (the so-called D-loop). (**b**) The telomeric 3′G-overhang can also form stable intramolecular and intermolecular four-stranded non-B DNA structures, called G-quadruplex structures

important role in avoiding end-to-end fusion by Homology-Directed Repair (HDR) and Non-Homologous End Joining (NHEJ) [24]. Besides preventing the telomeric region from being recognized as damaged DNA [27], the formation of t-loops is believed to be involved in the inhibition of telomeres elongation by telomerase [28] (see Sect. 7.3.3.1 for details).

The telomeric 3′G-overhang can also form stable intramolecular and intermolecular four-stranded non-B DNA structures, called the G-quadruplex structures (Fig. 7.1b). In humans, G-quadruplexes have been implicated in suppression of recombination, inhibition of telomerase activity, and telomere protection [27, 29]. Shelterin and CST complexes (see Sect. 7.3.3.1 for details) were shown to be involved in the formation and unfold of G-quadruplex structures at telomeres [30, 31].

7.3.2 The Importance of Telomeres' Replication for Genome Stability and Integrity

De novo DNA synthesis is fundamental to guarantee the survival of most species. Genetic material duplicates via a quality assurance mechanism that transfers on genetic information to make daughter cells identical to parental cells at each cell cycle. DNA replication is semi-conservative, meaning that each parental strand will serve as a template for the synthesis of a new complementary strand. A new DNA duplex is composed of one parental copy paired with one new DNA strand. After one round of replication, two identical copies of the original DNA molecules are synthesized. The process is catalyzed by DNA polymerases and complemented by additional proteins, such as helicases, primases, and ligases, resulting in the faithful copy of the genome.

In eukaryotes, a complex formed by DNA polymerase (Pol) alpha (Polα) and primase (PP) has both activities. Because DNA polymerases require a free 3′-OH group as the site for nucleotide addition in 5′ to 3′ direction, DNA replication differs among both strands. The leading strand is continuously duplicated in 5′ to 3′ direction by Polδ whereas in the lagging strand, PP firstly synthesizes primers of 7–12 mer ribonucleotides length that provide the 3′-OH free for Polε synthesis in a discontinued manner, processing small DNA sequences, called Okazaki's fragments. Since RNA primers are degraded after DNA polymerization, the replication of chromosome termini imposes a problem to DNA polymerases on the lagging strand. The last RNA primer is randomly positioned 70–100 nt from the 3′ end. Because conventional DNA polymerase has no free 3′-OH as the substrate for complete replication, removing the final RNA primer generates the 3′ G-overhangs in the newly synthesized telomere (G-strand). Yet, after DNA replication, nucleases Apollo and Exo I promote a resection at the leading strand telomeres (C-strand) so that 3′G-overhangs are generated at each chromosome end [32]. Slow and gradual shortening of the chromosome ends at every round of cell division was first described by

James Watson in 1972 and is known as the "end replication problem". This finding further increased the biological intrigue surrounding the study of eukaryotic telomeres [18].

A strategy used by eukaryotes to reverse this scenario draws on an enzyme specialized in elongating telomeres. The enzyme was named telomerase [17, 33], first described in 1985 in protozoa by Carol Greider and Elizabeth Blackburn (see Sect. 7.2). Telomerase has a conserved mechanism among eukaryotes and elongates telomeres using an intrinsic 451-nucleotide RNA molecule as a template [17, 33] (see Sect. 7.3.2.1). In addition to telomerase, there are alternative methods to elongate telomeres, consisting of homologous recombination pathways [27], which will be discussed later in this chapter (see Sect. 7.3.2.3).

Telomeres are considered the molecular clock of the cell [18, 27, 33, 34]. In mammalian, including humans, telomeres of non-proliferative somatic cells gradually erode after a certain number of cell divisions. In contrast, germline cells, totipotent cells and few somatic cells with proliferative capacity, including adult stem cells, maintain telomeres elongation by the action of telomerase [35]. As a consequence of inhibition of telomerase activity, with age, somatic cells will present progressive telomere shortening until they reach their replicative limit (replicative senescence), the Hayflick limit [9]. At the same time, many cell activities are suspended and even compromised when short telomeres become deprotected due to the insufficient amount of shelterin at chromosome ends, which can trigger a local DDR and induce apoptosis [15, 33]. Few cells escape senescence-associated antiproliferative checkpoints by deregulating tumor-suppression pathways and enter in crisis. In this state, most of the cells are eliminated by autophagic death [36]. Yet, cell survivors can reactivate the telomerase enzyme to maintain short telomeres and achieve replicative immortality, even presenting defective DDR and genomic aberrations. Telomerase reactivation is mostly due to mutations in the promoter of the *TERT* (Telomerase Reverse Transcriptase component) gene (see Sect. 7.6.1 for details) or upstream genomic rearrangements, present in as much as 90% of cancers, a strategy for unlimited replicative capacity or immortality [17]. Instead of reactivating telomerase, other cells use alternative mechanisms or ALT (Alternative Lengthening of Telomeres) (e.g. homologous recombination) to elongate telomeres and survive, both considered hallmarks of tumorigenesis [37].

7.3.2.1 The Telomerase Ribonucleoprotein Complex

Telomerase is a ribonucleoprotein (RNP) complex with reverse transcriptase function, able to elongate telomeres. The holoenzyme is minimally composed of the protein component TERT (encoded by the *TERT* gene) and the non-coding integral RNA TR expressed by the noncoding *TERC* gene (see Sect. 7.2), containing a 9-mer template sequence complementary to the telomeric repeat, copied by TERT during telomere synthesis [17, 33, 35]. The ribonucleoprotein complex also comprises several protein subunits, molecular chaperones, and co-chaperones such as heat shock protein 90 (HSP 90), NAF1, and accessory proteins of high importance for the conformation and assembly of the complex [38–40]. Among the main accessory proteins, the dyskerin complex (DKC), formed by dyskerin, NOP10, NHP2, and GAR1 proteins, plays a major role in folding and stabilization of the telomerase RNP complex. DKC binds a 3′-terminal region of the TR component that folds into two hairpin structures connected by a small nucleolar RNA (snoRNA) motif and a H/ACA box [41]. Telomere Cajal Body protein 1 (TCAB1) interacts with TR and is directly involved with the transportation of the mature telomerase RNP complex to telomeres by transferring the complex to the motile Cajal bodies [38, 39, 42] (Fig. 7.2a).

The catalytic subunit of TERT is highly conserved among organisms of different phylogenetic groups and preserves functional and structural similarities with conventional reverse transcriptases (RT). Similar to other DNA polymerases, its tertiary structure resembles a right hand (Fig. 7.3a). Human TERT protein comprises four structural domains, the TEN domain (Telomerase Essential N-terminal), the TRBD (TERT RNA Binding Domain), the RT (Reverse Transcriptase) domain, and CTE (C-Terminal Extension) domain (Fig. 7.3b) [43]. TEN and TRBD domains, located at the N-terminal region of the protein and CTE, are considered telomerase-specific and exclusive to each individual TERT protein, while the RT catalytic domain is the

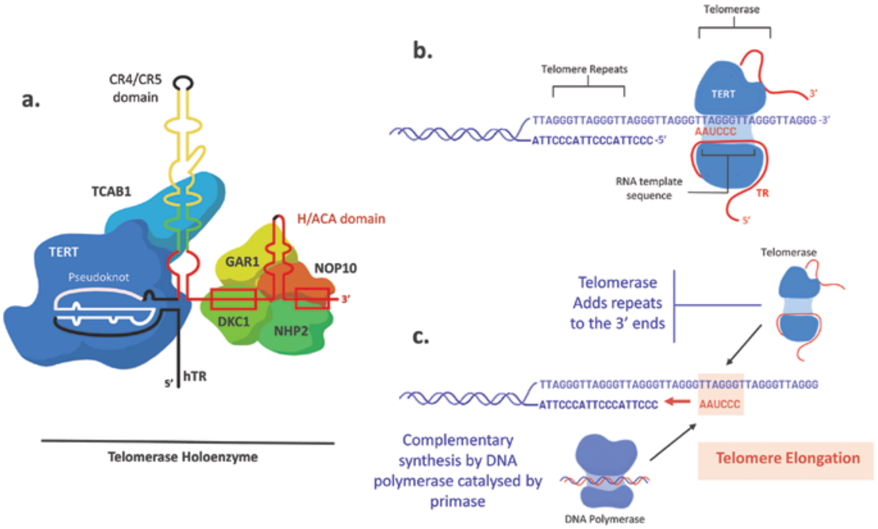


Fig. 7.2 Human telomerase ribonucleoprotein (RNP) complex and telomere replication. (**a**) The telomerase RNP complex showing the human telomerase holoenzyme: hTERT (Telomerase Reverse Transcriptase) and hTR (Telomerase RNA) components. hTR structure is formed by a pseudoknot containing the telomere template sequence, conserved regions 4 and 5 (CR4/CR5), and the H/ACA box domains (red boxes). Telomerase RNP accessory proteins are shown in colors: hTERT (blue), TCAB1 (light blue), DYSKERIN (DKC1 in green), NOP10 (orange), NHP2 (light green), GAR1 (yellow). (**b**) Telomerase RNP bound to telomeric DNA. (**c**) Telomere elongation by telomerase. Telomerase adds telomeric repeats at the chromosome 3′ends extending the G-rich strand and subsequently DNA polymerase alpha terminates telomere replication by C-strand fill-in synthesis

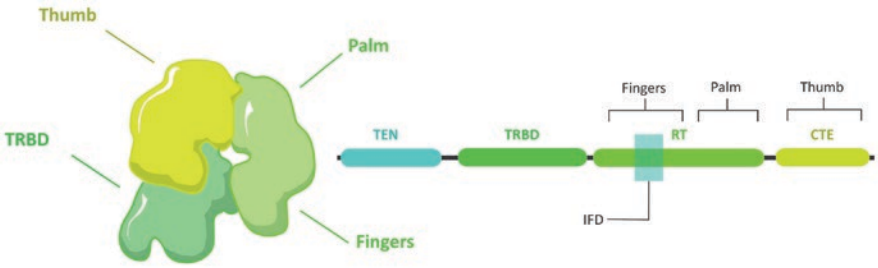


Fig. 7.3 Tridimensional structure of TERT catalytic subunit. (**a**) Illustrative image of the tertiary structure of hTERT, depicted as a "right hand" (analogy to fingers, palm and thumb) model. (**b**) Human TERT protein with four structural domains: the TEN domain (Telomerase Essential N-terminal), the TRBD (Telomerase RNA Binding Domain), RT (Reverse Transcriptase) domain with Insertion in Fingers Domain (IFD) and CTE (C-Terminal Extension) domain

most conserved and shared among different species and retrovirus [35, 44]. The TEN domain is divided into two sub-domains; the N-terminal-most TEN (TERT Essential N-terminal)/GQ/RID1 (RNA interaction domain-1) domain and TRBD/RID2 (RNA Interaction Domain-2) domain, separated by a nonconserved sequence linker. TEN/GQ/RID1 domain is implicated in DNA substrate recognition and involved in telomere maintenance and telomerase activity processivity. GQ binds the shelterin component TPP1 during telomerase recruitment for telomeres elongation. The TRBD domain contains the CP and T-FLY motifs, and in addition to bind TER, it is also involved in regulating telomere synthesis. The catalytic center of the enzyme contains the reverse transcriptase domain, with several conserved motifs. Motifs 1 and 2 comprise the region known as "finger", contains and includes the "insertion in fingers domain" (IFD), which is related to the stabilization of interactions typical of hTERT and associates with the G-rich telomeric DNA. Motifs A to E that refer to the "palm" region concentrates the hTERT catalytic core. The CTE domain, representing the "thumb", varies in sequence between organisms of different species, suggesting that it has specific functions [41, 43–45] (Fig. 7.3a).

Recently no-telomeric roles have also been imputed to hTERT. These include the involvement in DDR regulation, cell growth and proliferation, cell cycle kinetics, and the protection of mitochondria integrity and mitochondrial DNA (mtDNA) damage from oxidative stress [46]. Understanding how these non-canonical functions impact its primordial function for telomere length maintenance remains to be determined.

In most eukaryotes, including humans, the telomerase TR component is a polyA+ noncoding RNA transcribed by RNA polymerase II. It diverges in sequence and length among organisms of different species and provides the structural scaffold for assembling the telomerase RNP complex [28, 47]. The human 451-nucleotide non-coding TR component of the telomerase is a small RNA that belongs to the small Cajal body RNAs (scaRNAs) and snoRNAs family [47, 48].

HTR presents four structural and functional domains (Fig. 7.2a). At the 5′ region, a RNA pseudoknot domain contains the template sequence 5′-AAU**CCCAAU**C-3′ (the sequence complementary to the human hexanucleotide telomeric repeat unit in bold) copied by hTERT during telomere elongation (Fig. 7.2b) [41, 44]. The Stem Terminus Element (STE) comprises the conserved regions 4 and 5 (CR4/CR5) and the template boundary element (TBE), which respectively binds TCAB1 and hTERT via the TRDB domain. TBE limits the RNA access to the RT active site while the CR4/CR5 facilitates the association with hTERT. Additionally, during the RNP assembly, TBE helps to lead the telomerase components to Cajal bodies in the nucleus. The H/ACA box domain, located in the 3 ‘region of TR, contains the H-box (a single-strand hinge) followed by a stem-loop and the sequence ACA, which together with the conserved region 7, are important for hTR biogenesis, telomerase activity, and hTR stability. The H/ACA domain also contains a conserved four-nucleotide motif called the CAB box, which binds the accessory proteins DKC1, GAR1, NOP10, NHP2, NAF1 to form the RNP complex. Once assembled, telomerase can be recruited to telomeres through the TPP1 protein component of the shelterin complex [39, 41].

Because hTR is constitutively expressed independently on the cell type, it can be detected in most somatic cells that do not present telomerase activity, since the transcription of the human *TERT* gene is repressed [39]. Constitutive expression of hTR is intriguing and reinforces that it may participate in other non-telomeric cellular biological processes. It was recently shown that hTR transcripts encode a 121 polypeptide (hTERP) involved in cell protective pathways such as autophagy and apoptosis [49]. As for hTERT, these alternative hTR functions should be deeply investigated and better contextualized in terms of cell homeostasis. Besides, hTR mutations are also involved with telomere dysfunctions currently found in patients with telomeropathies [50, 51], as presented later in this chapter (see Sect. 7.5 for details).

7.3.2.2 Elongation of Telomeres by Telomerase

As mentioned before, incomplete lagging strand synthesis generates 3′G-overhangs in each chromosome ends, which can induce progressive telomere shortening after each round of cell duplication [52, 53]. Human telomerase uses the 3′G-overhang as substrate and adds TTAGGG repeats by copying the hTR template sequence in proliferative cells. At the same time, DNA polα-primase completes the C-strand fill-in synthesis. However, telomere elongation depends on the coordinated action of the shelterin components TPP1-POT1 and the CST complex, which respectively associate with telomerase RNP and DNA polα enabling telomeres elongation and C-strand synthesis [32] (see Sect. 7.3.3 for details).

Elongation of telomeres by telomerase is cyclic and initiates with the complementary interaction between the telomeric 3′G-overhang and the hTR template sequence that corresponds to one telomeric repeat. Additional interactions of upstream telomeric DNA with the hTERT TEN domain are also important. Subsequently, the hTERT catalytic site adds nucleotides onto the 3′ end of the telomeric DNA by copying part of the hTR template sequence. During synthesis, hTERT uses its characteristic processivity across the template. First, telomerase adds several nucleotides to telomeres (type I processivity). Then, hTR translocates to reposition the 3′ end of the template at the 3′ end of the newly synthesized repeat. Additional processivity can now occur (type II processivity), meaning that a number of telomeric repeats can be added by telomerase in a single interaction event with telomeres (Fig. 7.2b, c). The active telomerase cycle can now be repeated many times without fully releasing telomeric DNA [17].

7.3.2.3 Alternative Lengthening of Telomeres

Tumor or immortalized cells that cannot reactivate telomerase to elongate telomeres use alternative mechanisms referred to as Alternative Lengthening of Telomeres (ALT). This mechanism is mainly based on Homologous Recombination (HR) among telomeric DNA sequences [54].

Cells that employ ALT (ALT cells) present many unique characteristics such as long and heterogeneous telomeres and the presence of non-canonical telomeric repeats called Telomere Variant Repeats (TVRs e.g., TCAGGG) [55]. Also, ALT-positive cells display a permissive chromatin state, extrachromosomal telomeric DNA in the form of T-circles (double-stranded telomeric circles), and single-stranded C-rich or G-rich circles [56, 57]. Curiously, TVRs do not bind the shelterin proteins TRF1 and TRF2 (Telomeric Repeat-binding Factors 1 and 2; see Sect. 7.3.3.1). They associate with nuclear receptors that help to spatially approximate telomeres and to promote telomere-telomere recombination via association with proteins that mediate break-induced telomere elongation [57–59].

An additional characteristic of ALT cells is the clustering and encapsulation of telomeric DNA within promyelocytic leukemia nuclear bodies (PML-NBs), forming ALT-associated PML bodies or simply APBs [60]. APBs are mainly composed of proteins such as PML and SP100, several telomeric proteins (e.g., TRF1, TRF2) and proteins that participate in HDR (e.g., ATR kinase, RAD51, and RAD52, among others). Their formation is dynamic and driven by SUMOylation. It occurs in a cell-cycle dependent manner, principally in G2 phase, stimulated mainly by DSB DNA at telomeres and replication stress [61, 62]. Therefore, APBs are considered sites of recombination, and probably of telomere synthesis and generation of extrachromosomal telomeric circles. Recent reports evidence the fact that ALT cells are highly sensitive to inhibition or depletion of HDR factors, which could be a useful tool for future therapy against cancer [63].

7.3.3 *Telomere Length Regulation*

7.3.3.1 The Intercrossed Actions Between Shelterin and the CST Complex

Telomeres are known to promote genome stability and integrity by preventing DNA damage signaling and the activation of local DDR pathways, which would recognize chromosome termini as DSB. These abilities hinge on, among other factors, the shelterin complex and its functions (Fig. 7.4). The shelterin complex is a hexameric

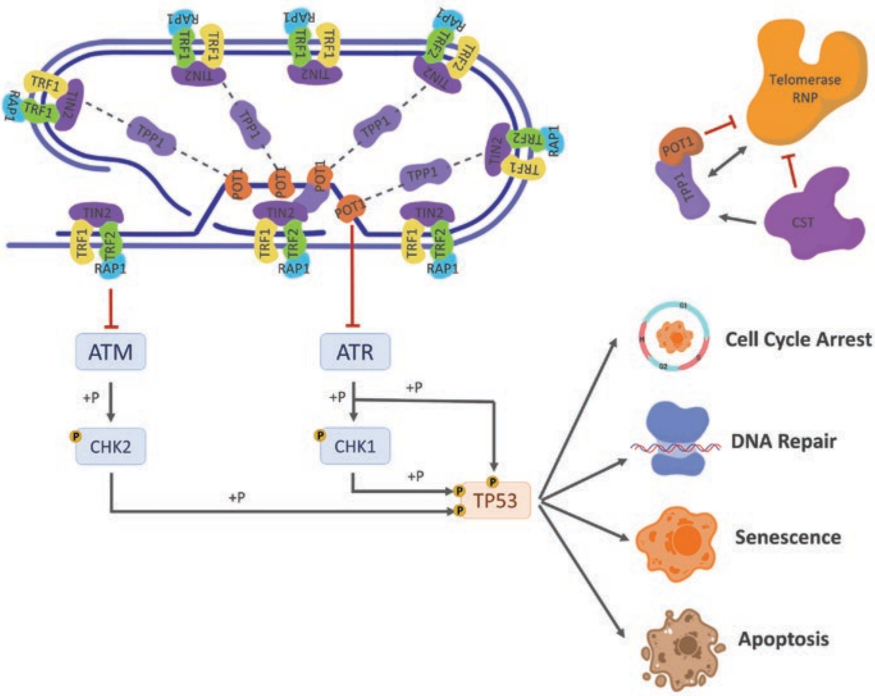


Fig. 7.4 Shelterin complex protects telomeres from DNA damage response. A cartoon showing the six shelterin protein components at human telomeres. Depiction of TRF2 and POT1 inhibition of the ATM and ATR kinase pathway can induce through TP53 different fates: cell cycle arrest for DNA repair, senescence or apoptosis. TPP1/POT1 can play a dual role at telomeres by recruiting or inhibiting telomerase access. The possible cellular events that result from telomere deprotection are also illustrated. The binding area of TRF1 and TRF2 is a tightly packed chromatin. TIN2 also interacts with TRF2, but its relevance remains unknown. The cartoon is drawn without this linkage to illustrate the formation of subcomplexes. The CST complex can also protect telomeres by interacting with TPP1/POT1 inhibiting telomerase access to telomeres, and recruit DNA polymerase alpha for the C-strand fill-in synthesis. Unidirectional arrows indicate activation, the bidirectional arrows indicate enzyme recruitment activity, the red lines indicate inhibition and the dashed lines indicate physical association

telomere-specific protein complex, constitutively expressed in cells. Shelterin complex proteins and other factors collaborate in the regulation of telomerase activity. The complex also participates in diverse activities, from inhibiting the activation of DDR by ATM, ATR kinases, and PARP1 (poly (ADP-ribose) polymerase 1) through blocking DSB repair by classical NHEJ (c-NHEJ), alternative (alt-NHEJ) and HDR, to preventing harmful telomere resection [24, 64]. Shelterin complex in humans and some other mammalian species comprises six major interacting proteins (POT1, protection of telomeres 1; TRF1 and TRF2; TIN2, TRF1- and TRF2-interacting nuclear protein 2; TPP1, TIN2 and POT1 interacting protein 1; and RAP1, repressor/activator protein 1) that bind both double and single-strand telomeric DNA [27] (Fig. 7.4).

The assembly of protein subunits belonging to the shelterin complex must be sufficiently functional to offer telomere protection [65]. TRF1 and TRF2, although structurally different, bind as homodimers to the duplex region of the telomere DNA using an MYB/SANT DNA-binding domain [19, 66]. TRF2 independently recruits RAP1, which does not bind DNA, through an interaction with the RAP1 RCT domain, and this association represses aberrant HR at telomeres [67]. The binding domains present on TRF1 and TRF2 facilitate their interaction with accessory proteins and engage them in various activities at the telomeres. Most of these accessory proteins function inside the nucleus, while a few others operate in the cytoplasm. Apollo and tankyrase are typical examples of such accessory proteins. Tankyrase 1 and 2, for instance, inhibit the binding of TRF1 to the duplex telomere region as it post-translationally modifies TRF1 [68]. TRF1 and TRF2 also bind TIN2 using different interaction surfaces, and TIN2 bridges the TRF1/TRF2-RAP1 subcomplex to the TPP1/POT1 heterodimer, ensuring concurrent connectivity amongst the subunits [69, 70]. TPP1 does not interact with DNA but is multifaceted since it associates with the hTERT TEN domain via a patch of surface aminoacids (TEL-patch), being important for telomerase recruiting to telomeres and enzyme processivity [71]. Together with POT1, TPP1 plays a protective role at telomeres by repressing local DDR and chromosome fusions [26, 72]. POT1, in its turn, binds the G-rich single-stranded telomeric DNA (3′G-overhang) using two structural OB (oligonucleotide/oligosaccharide)-fold domains [73]. POT1 binds near the junction between the double- and single-strand DNA of the telomere, which gives it an advantage in performing its role in protecting telomeres and interacting with TPP1 via intermolecular linkage (Fig. 7.4). POT1 preference for the G-rich single-stranded telomeric DNA does not impair the CST complex (see below) to compete for binding to the same substrate (the G-rich single-stranded telomeric DNA). However, POT1 plausibly relies on its relationship with TPP1, which is already present at the site, to get ahead of the CST complex [74].

It is clear now that a combination of four of the subunits of shelterin containing TRF1, TIN2, TPP1, and POT1, in a 2:1:1:1 stoichiometric composition, is sufficient for the shelterin complex to assemble at the telomere. The four polypeptides are capable of binding DNA in double and single-strand forms, pointing to shelterin capabilities in forming subcomplexes along the entire length of the telomere (Fig. 7.4) [64, 65, 75, 76].

Besides shelterin, the CST (CTC1, STN1 and TEN1 proteins; see below) complex also contributes to telomere maintenance and preferentially binds the G-rich single-stranded telomeric DNA. CST was first described in budding yeast, and CST-like complexes were subsequently reported in many other eukaryotes. A common feature shared among all CST complexes described so far is their structural similarities with the components of the heterotrimeric replication protein A (RPA) complex since all of them contain at least one OB-fold domain [77–80]. RPA binds non-specifically single-stranded DNA and is involved in many DNA metabolism pathways, including DNA replication and repair [81].

Human CST (hCST) is composed of three proteins; CTC1 (Conserved Telomere maintenance Component 1), STN1 (Suppressor of CDC Thirteen homolog) and TEN1 (Telomere length regulation protein TEN1 homolog). Its main role at telomeres is limiting telomerase action and promoting C-strand fill-in synthesis. More specifically, hCST inhibits telomerase activity through primer sequestration and physical interaction with POT1/TTP1 regulating DNA polα access to telomeres. Both events are coordinated with the shelterin complex [82].

Human CTC1 is the large CST subunit and shows high affinity and specificity to the telomeric single-stranded DNA. Though human STN1 binds telomeric DNA with low affinity and no specificity, human TEN1 does not bind DNA [83]. CTC1 and STN1 are also involved in many protein-protein interactions facilitating telomerase and DNA polα recruitment to telomeres [80]. Natural mutations in the *CTC1* and *STN1* genes result in rare genetic, clinically overlapping, human disorders such as Coats plus syndrome and dyskeratosis congenita owing to dysfunctional telomeres [13, 84–86] (see Sect. 7.5 for more details). Human STN1 and TEN1 can also form a CST-subcomplex with roles in DNA replication, meaning that they can play extratelomeric functions [87].

7.3.3.2 Telomere Transcription and the Importance of Terra in Telomere Maintenance

Telomeric repeat-containing RNAs (TERRA) are long non-coding RNAs (lncRNA; see Chap. 5) described more than two decades ago in several eukaryotes [88–90]. Since then, there is crescent evidence about their importance in different biological processes with emerging roles in telomere maintenance and genome stability [91–93]. TERRA stands out as the most studied RNA of the telomeric transcriptome, although details about their mechanism of action remain to be elucidated [93–95].

TERRA is transcribed from the telomeric C-rich strand by RNA polymerase II and its transcription is driven by subtelomeric CpG island promoters [89, 90, 96–98]. In humans, TERRA sequence expresses repeat-containing transcripts 100 bp to 9Kb in length, consisting predominantly of subtelomeric sequences and 5′UUAGGG 3′ repeats ranging from 100 to 400 bp in the 3′ end [88, 90, 99]. Mature TERRA transcripts contain a 7-methylguanosine cap and in humans, approximately 7% of them are polyadenylated [100]. However, telomeric chromatin represses TERRA transcription while it is upregulated at dysfunctional telomeres. TERRA upregulation is induced by the removal of TRF2, promoting the accumulation of histone 3 lysine 9 trimethylation (H3K9me3; see Chap. 4) [101–103] and telomerase inhibition [104]. Curiously, TERRA transcription is also upregulated in the Immunodeficiency, Centromeric region instability and Facial anomalies syndrome (ICF) type I syndrome, a very rare autosomal human disorder in part due to hypomorphic mutations in the *DNMT3B* (DNA Methyltransferase 3B) gene, responsible to methylate DNA at subtelomeric regions and extra repetitive DNA. Carrier patient cells also show accelerated telomere shortening and prematurely enter replicative senescence [105]. It was recently demonstrated that in these patients, elevated TERRA results in the formation of aberrant RNA:DNA hybrids (see the description of TERRA R-loops below), triggering DNA damage and telomere instability [106].

TERRA expression downregulation induces DDR at telomeres and formation of "Telomere dysfunction-Induced Foci" (TIFs) [107, 108]. In agreement, knockout of TERRA locus from telomere 20q in human osteosarcoma cell line U2OS increased DNA damage at chromosome ends, telomere fusions and telomere shortening [109], confirming a role for TERRA in telomere maintenance. TERRA abundance depends on the cell cycle phase (in humans, it is more abundant in G1 than S phase), cell development state, telomere length and telomerase activity. For example, ALT cells express high TERRA levels compared to telomerase-positive cells, which can be explained by less compact chromatin at ALT cells telomere [99, 110]. Therefore, the interaction of TERRA transcripts with heterochromatin marks can affect its regulation and telomere maintenance. The impact in telomere maintenance can happen through TERRA transcripts' interactions with telomeric proteins, such as the shelterin component TRF2, and other proteins present at telomeres (hnRNPa1, Suv39h1 and ORC1, origin replication complex 1). In this context, TERRA acts as a scaffold recruiting other proteins to the chromosome ends (e.g., histone-modifying enzymes and chromatin remodeling complexes) [102, 111–115].

TERRA can also form both RNA:DNA hybrids, named TERRA R-loops and TERRA G-quadruplex. TERRA R-loops structures can influence heterochromatin formation at telomeres. In high levels, they increase the formation of DNA DSB, promoting HDR at telomeres, affecting the integrity of the subtelomeric and telomeric regions [93, 108, 116, 117]. TERRA G-quadruplex acts as a binding target for telomere-binding proteins, such as TRF2, and promotes histone H4 trimethylation. Thus, it is conceivable that TERRA transcription regulation is not only associated with telomere integrity but may also be implicated with different chromatin events, regulation of gene expression, senescence, and aging [117]. Notably, TERRA seems also to play extratelomeric roles. Recently it was shown that only a subset of TERRA transcripts accumulates at telomeres as they transiently localize at chromosome ends and can also be found interacting with extratelomeric loci. At these extratelomeric loci TERRA regulates the transcription of subtelomeric and internal chromosome genes [115, 118]. Whether TERRA can regulate epigenetic signatures at subtelomeric and extratelomeric loci should be demonstrated [93, 119].

7.4 Telomeres and Biological Age

The process of aging can be translated into a time-dependent gradual and multifactorial alteration on the homeostatic status of an individual, with the most significant physiological result being the loss of functionality at multiple biological levels. In humans, functional decline resulting from aging is a risk factor for several pathological conditions, such as cardiovascular diseases, degenerative diseases, cancer, diabetes, or obesity. From a pathophysiological perspective, biological aging means the accumulation of DNA damage, oxidative stress, loss of proteolysis, mitochondrial dysfunction, cellular and tissue senescence, and telomeric friction [120]. Central to many of the these-mentioned "hallmarks of aging" as denominated by López-Otín (2013), telomere shortening is a relevant element in human aging. Therefore, the natural process of telomere shortening can be a potential indicator of the biological pace or rhythm of aging, reflecting the proliferative history of the cells [121, 122]. In this sense, we can extend the definition of telomere length to a metric of biological aging concerning to chronological aging. Suggesting that this definition sum to the concept of aging, as all cellular events that the organism experienced. Supporting this perception, in humans, several epidemiological studies, as well as literature reviews, have shown that there is a negative correlation between an individual's age and their telomeric length [123–125]. For example, data published in a study with 981 participants, ranging from 45 to 84 years, revealed a negative correlation between the chronological age and telomere length [126–129]. Corroborating these data, important results published from the Resource for Genetic Epidemiology Research on Aging (GERA) cohort, with 105,539 participants, confirmed the predicted age-dependent decline in telomere length [127]. Thus, these results provide us with strong evidence that telomere length can serve as a marker of the individual's biological age (Fig. 7.5). However, it is important to note that characteristics such as ethnicity and sex significantly modulate this relationship and cannot be disregarded [130].

7.4.1 Environmental Determinants of Telomere Shortening

Humans telomeres measures between 10 to 15 Kb at birth [127, 128]. As explained in Sect. 7.3.2, these sequences are shortened during mitosis at an average rate of approximately 65 bp per year since the DNA replication machinery cannot replicate the last fragment at the end of the lagging strand [131–135]. However, it is essential to note that this rate is highly variable and subject to endogenous and external modulations, such as the individual's physiological condition and behavior, the environment where he lives, or the pathologies that may be associated. Therefore, it is clinically important to point out elements capable of determining this modulation. In this section, we will focus on the extrinsic factors responsible for this regulation.

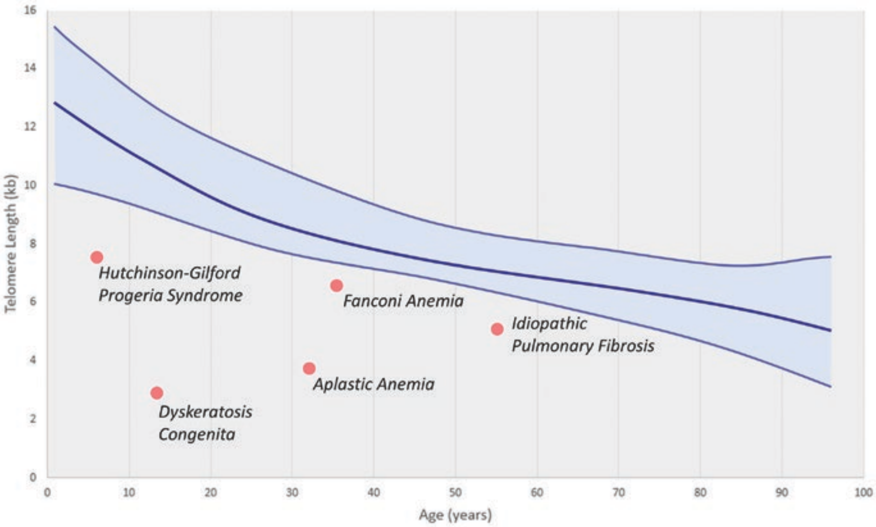


Fig. 7.5 Leukocyte telomere length by qPCR. Illustrative image of mean telomere length (blue thick and thin lines represent central, minium and maximum values) of peripheral blood leukocytes from healthy volunteers measured by quantitative polymerase chain reaction (qPCR) from 0 (newborn) to 100 years. Representative values from telomeropathies (Hutchinson-Gilford, dyskeratosis congenita, aplastic anemia, Fanconi anemia and idiopathic pulmonary fibrosis) are shown as red dots

In this sense, it is recognized that a wide range of behavioral habits compromises telomere length. Among them, smoking [123, 136–138] and alcohol consumption [139, 140] are factors classically associated with this phenomenon. Furthermore, individuals subject to continuous environmental and psychological stress may experience accelerated shortening of telomeres. It has recently been shown that the telomere length of newborns correlates with the socio-economic status of their parents [141, 142]. That is, social vulnerability, an intrinsically environmental factor, has the potential to alter the telomere length of individuals and possibly an implication for reduction in longevity. This evidence seems to extend to other domains of influence and demonstrate an intergenerational modulatory potential of the factors that are harmful to telomeric length. A study conducted with 4,935 individuals concluded that men and children who were exposed to risky parental environments had reduced telomeres. Also, in women, parental substance abuse appears to affect telomere length [143].

In contrast, some factors can cause the opposite effect and reduce the speed of telomere loss. In this sense, epidemiological evidence demonstrates important inverse correlations between regular physical activity and telomere erosion [144]. Recently published evidence in a study comprising 2,401 twin subjects demonstrated a positive correlation between leukocyte telomere length and physical activity. Interestingly, this correlation remained, even when factors such as smoking, body mass index (BMI) or socioeconomic status were considered [129, 145, 146].

These results align with the benefits of exercise in aging and general health described in the literature. For example, a recently published study, comprising data collected from 661,137 individuals, demonstrated that those who practice leisure-time physical activities, even below the recommended by the Physical Activity Guidelines for Americans, have a 20% lower mortality risk [129].

Many mechanisms have been proposed over the years to explain from a molecular point of view the relationship between these elements and the modulation of telomere length. Among them, changes in the dynamics of oxidative stress, the inflammatory process, or the telomerase reactivation are frequent candidates [147]. The impact of these factors on telomere erosion will be discussed below.

7.4.2 Intrinsic Factors for Telomere Shortening

Central to many of the conditions associated with aging, homeostatic dysfunction, and chronic inflammation have a prominent and widely described role in the literature. In this sense, telomeric friction is located at the apex of the relationship between chronic inflammatory diseases and aging, and is a major factor in establishing the premature aging phenotype. A mechanism commonly altered in an inflammatory disease's pathological conditions is the dynamics of production of reactive oxygen species (ROS). In this sense, oxidative stress is a relevant element in telomeric modulation, being responsible for a significant increase in telomere shortening rate. That is, cells exposed to extrinsic pro-oxidative conditions, such as tissue hypoxia, or even intrinsic conditions, such as mitochondrial dysfunction, tend to have accelerated reduction in telomere length [126, 148–153]. This relationship is corroborated by a recent study, comprising a cohort of more than 500 individuals that found a correlation between telomere length and the expression of genes involved in the ROS production pathways [149, 154, 155]. Besides, another study conducted with the cohort from the Framingham Heart Study demonstrated a relationship between markers of systemic oxidative stress and telomere erosion [148]. Furthermore, oxidative stress production directly impacts the rate of DNA damage to which cells are subjected [126]. In this sense, persistent DNA damage from the chronic inflammatory processes can lead to growth arrest, replicative senescence, or apoptosis, correlating with the uncapping of the telomeres and a general loss of telomere length [149, 151] (Fig. 7.4). These phenomena can mediate the relationship between some pathological conditions and the observed telomere shortening.

7.5 Telomere Dysfunction in Human Diseases

Telomere dysfunction due to damaged maintenance of the telomere machinery contributes to the inherited telomere syndrome spectrum. While clinically diverse regarding the age of onset and symptoms, these monogenic genetic diseases share similar underlying molecular mechanisms and have overlapping phenotypes connected to telomere biology-related genes, causally linked to critically short telomeres [124, 125]. Short telomeres are commonly associated with some aging-related diseases, such as cancer and a spectrum of diseases caused by mutations in genes for some telomerase RNP components collectively known as telomeropathies (Fig. 7.5) [27, 33, 51, 156, 157]. Additional factors may cause telomere shortening, such as the telomeric transcriptome's disbalance, mainly the upregulation of TERRA transcription (see Sect. 7.3.3.2 for details), inflammation and oxidative stress [33].

Dyskeratosis congenita (DC) is a complex pathology with different trigger mechanisms. Clinical manifestations that provide the basis for DC definition are characterized by the mucocutaneous triad of abnormal reticulated skin pigmentation, nail dystrophy and mucosal leukoplakia [13]. These clinical manifestations are accompanied by inherited bone marrow failure [158], the leading cause of death in DC patients. Due to bone marrow failure, individuals with DC have an increased risk of developing several malignancies, such as myeloid leukemia and squamous cell carcinomas [159]. DC patient cells present short telomeres and the disease is thus considered an example of the spectrum of primary telomeropathies (telomere-syndrome-diseases). The first evidence that causally related shorter telomeres and DC was the description of a rare DNA variant in the X-linked *DKC1* gene, which codes for a critical protein of the telomerase holoenzyme, dyskerin. The *DKC1* mutation causes an X-linked recessive form of DC and is associated with reduced TR expression and consequently reduced telomerase activity, a key feature contributing to telomere shortening [160]. Further studies with genome-wide linkage analysis and gene sequencing led to the identification of other targets in DC, such as *TINF2* and or telomerase-related gene mutations which cause autosomal dominant (*TERT*, *TERC*) or recessive (*TCAB1*, *NOP10*, and *NHP2*) forms of the disease, associated with anticipation and more severe clinical manifestations [48].

Other extremely rare monogenic Mendelian diseases show pathobiology and clinical overlapping with DC and are recognized as DC variants with markedly shortened telomeres and lifespan. For example, Hoyeraal-Hreidarsson syndrome (HH) presents as a severe type of DC with mutations in the *DKC1* gene. Patients with HH syndrome present cerebellar hypoplasia and microcephaly and critically short telomeres [161, 162]. Revesz syndrome is caused by *TINF2* mutations, a rare pathology with clinical presentation similar to HH but severe retinopathy [163]. Finally, patients with Coat plus syndrome with mutation in the *CTC1* gene present intracranial calcification in addition to similar symptoms as those observed in HH and Revesz syndromes [164].

Aplastic anemia (AA) is a rare nonmalignant hematologic disorder characterized by injured and impaired bone marrow, features that overlap with DC [165]. Due to progenitor cell impairment, a representative subset of patients with AA presents short telomeres in cells from the immune system compartment that correlate with disease severity. Critically shorter telomeres observed in hematopoietic stem cells are associated with *TERT*, *TERC* and *DKC1* gene mutations, a feature shared with DC [166]. Short-range pathogenic DNA variants in the *TRF1* and *TRF2* genes encoding components of the shelterin complex have been identified as possible candidates for disease risk [167]. Additionally, a biallelic mutation in *RTEL1* was identified in about 1% of AA cases [168].

Idiopathic pulmonary fibrosis (IPF) is a rare and distinct type of chronic lung fibrosis and the most common manifestation of telomere syndrome in adulthood [169]. The etiology and progression of IPF are highly variable. The disease's pathophysiology is associated with age-dependent compromise of lung function due to recurrent injury in alveolar epithelial cells [170, 171]. Some patients show slow progress while others show periods of stability interspersed with acute deterioration, rapid respiratory function decline and organ failure. Genetic variants in genes related to the telomere biology (*TERC*, *TERT*, *PARN*,[1] *RTEL*,[2] [172] and *STN1*) contribute to the augmented risk for disease development [173–175]. Another mutation in the *TINF2* gene encoding the shelterin protein TIN2 is described as a factor contributing to a small subset of familial cases [176]. In the absence of familial gene mutations, short telomeres are an important risk factor for the onset of a heterogeneous type of IPF [177]. Frequently, individuals affected by one disease have subclinical manifestations in other organs. Individuals with telomerase complex gene mutations leading to IPF can also have bone marrow failure and liver disease. The co-occurrence of AA and IPF, or the presence of myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML) are recurrently found in DC and IPF patients [178]. Simultaneous manifestations of apparent different diseases, now associated by an underlying molecular mechanism of telomere dysfunction, suggest that IPF, AA and liver cirrhosis can be considered different manifestations of a single wide spectrum disorder [179].

Hutchinson-Gilford Progeria Syndrome (HGPS) is a known premature aging syndrome caused by LAMIN A/C (*LMNA*) gene mutation leading to progerin protein accumulation at the nuclear membrane, promoting a loss in the regular architecture of the chromatin [50, 179]. Progerin expression promotes the accumulation of senescent cells through the activation of telomere-specific DNA damage leading to telomere dysfunction [180, 181]. Additional studies show that progerin may disrupt the interaction between telomeric proteins and the nuclear lamina through RAP1, TRF1 and TRF2, triggering telomere dysregulation and shortening [182, 183]. The diseases previously described grouped by mutations in telomere maintenance-related genes are denominated as primary telomeropathies.

Secondary telomeropathies group diseases due to mutations in genes involved in DDR that compromise tissue repair, disrupting senescent cell clearance, and indirectly promoting premature telomere attrition. While symptoms and disease manifestations remain, the senescent cell phenotype observed in these diseases can be rescued by telomere elongation through telomerase activity [184–186].

A hallmark of monogenic human telomere syndromes is genetic anticipation. It seems that short telomeres may contribute as a deterministic factor and cause different diseases with similar phenotypes. In some cases, parental telomere

[1] PARN: Poly(A)-Specific Ribonuclease.

[2] RTEL: Regulator Of Telomere Elongation Helicase 1.

shortening may anticipate the age of onset and accelerate telomeropathies' signs and symptoms in offspring, increasing disease severity and further anticipation in later generations, resulting in early death. In this line, *TERT*/*TERC* mutations might promote adult-onset such as in IPF and AA, whereas younger generations may develop DC-related diseases. These observations, supported by studies carried out in mice, provide evidence that telomere shortening is the major mechanistic cause of these syndromes so that non-carrier offspring can manifest such diseases by inheriting shorter telomeres from a parent [182, 187, 188].

Dysfunctional or short telomeres may have a contribution to many common age-related diseases. According to these findings, high genetic risk for coronary arterial disorders can associate with single-nucleotide polymorphisms (SNPs) in genes related to telomere biology [13, 189]. On the other hand, accelerated telomere shortening has been observed in peripheral blood leukocytes as a predictor variable associated with numerous metabolic chronic and inflammatory diseases [190, 191], such as cardiovascular diseases [192, 193], diabetes [194], ulcerative colitis [150, 151], obesity [195] and chronic inflammation-related comorbidities [196–198]. Several pulmonary diseases [199–201], such as chronic obstructive pulmonary disease (COPD) [202] and severe asthma [203], are also associated with telomere attrition. These chronic metabolic disorders share important physiological features closely related to chronic pro-oxidant and pro-inflammatory milieu that may contribute to the persistence of such diseases.

Short telomeres are important in neurodegenerative diseases [204, 205], and this is sustained by SNPs in *TERT* and *STN1* (OBFC1 associated with augmented risk for developing Alzheimer's disease (AD) [206–208]. Mental disorders and chronic systemic low-grade inflammation are also described in association with shorter telomeres [209] affecting brain volume and memory performance [210, 211]. Interestingly, as aforementioned, environment may play an important role during disease persistence. This is particularly important in several mental illnesses, as observed in schizophrenia, in which even healthy siblings of affected subjects display shorter telomeres [212].

Finally, telomere shortening-related dynamics can be detected even in early childhood when evaluating diseases [213], psychological disorders [214] or lifestyle factors and conditions [215, 216]. Nonetheless, these observations usually are associated with worst clinical outcome in adulthood [194, 217].

7.6 Telomeres as Targets for Disease Management

7.6.1 Telomerase Inhibition in Cancer

Strategies for exploiting telomerase inhibitors are still in preclinical phase. Inhibitors of telomerase, immunotherapies and targeting the *TERT* gene expression driven by its promoter mutations have been extensively studied in the field of cancer. Physiological *TERT* inhibition results in telomere attrition upon division. Critical short telomeres induce DDR, cell cycle arrest leading to DNA repair, senescence or apoptosis via the TP53/retinoblastoma protein suppressor pathway. Telomere attrition acts as a tumor barrier for immortality. Replication beyond a critical point leads to instability and eventually elicits telomere crisis. Extensive genome instability primes cell death. Because heterogeneity is a feature of tumorigenic cells, few clones escape that crisis and reactivate telomerase for telomere maintenance [218]. Unlike human adult somatic cells in which telomerase activity is virtually silenced, telomerase reactivation for telomere maintenance in transformed cells is nearly a universal hallmark across different cancer types [219]. Expression (~75%) and activity (~90%) of TERT is characteristic of cancer cells that overcome the senescent replicative state [220]. For that, TERT is an attractive candidate for clinical therapies. Cryo-electron microscopy improvements in structural resolution (7–8 Å) [221] will facilitate targeted compounds' design against TERT.

Human TERT polypeptide (1,132 amino acids) is processed by the proteasome cellular machinery and presented by cancer cells in the context of MHC I molecules. Additional peptides that bind MHC class II alleles have been discovered, suggesting that cancer cells can present TERT peptides to CD4+ and CD8+ helper and cytotoxic T lymphocytes. The finding that TERT can be processed as tumor-associated antigens [222] and induce antigenic and immunogenic responses led to the development of different strategies or immunotherapies for targeting telomerase. TERT peptide vaccines, adoptive cell transfer and oncolytic virotherapy have been tested in preclinical settings. For example, most vaccine trials have used synthetic hTERT peptides in phase I/I-II trials [223]. Results are not encouraging with mild TERT-specific T-cell responses, minimal effect on tumor size, and temporary disease stabilization, irrespective of cancer type. A phase III TERT vaccine called GV1001 in patients with advanced pancreatic cancer failed to demonstrate any survival advantage over chemotherapy [224]. In light of these findings, TERT peptide vaccines may be boosted in combination with immune checkpoint blockade [225]. Additional strategies that explored TERT-positive expression in transformed cells used a viral approach. An oncolytic adenovirus called Telomelysin (OBP-301; Oncolys BioPharma Inc., Tokyo, Japan) that replicates in tumoral cells via expression of E1 protein under the regulation of *hTERT* promoter was tested in phase I clinical trial [226]. However, direct intratumor injection had limited immune response. In light of these results, pharmacological inhibitors have been developed or repurposed. Finally, the telomerase activity has been blunted by the incorporation of called "uncapping agents". For instance, 6-thio-2′-deoxyguanosine (6-thio-dG) triphosphate impedes the shelterin complex's binding to telomeric DNA, activates the DDR, and induces telomere dysfunction and cell death in telomerase-expressing cells [227]. 6-thio-dG has proven effective in different clinical settings (non small cell lung carcinoma (NSCLC), medulloblastoma, or melanoma xenografts).

Imetelstat (Geron, CA, USA) is a telomerase inhibitor based on an oligonucleotide sequence complementary to the RNA template sequence from TR. Its mechanism relies on selective competition and inhibition of telomerase. Active proliferative tumor cells undergo telomere shortening, induction of DDR, and cell death. NSCLC phase II trial had no overall survival but clinical improvement in a portion of patients with short telomeres [228]. This observation supported repurposing Imetelstat for myelofibrosis, although the molecular mechanism is not understood [229].

G-quadruplexes (G-4s) are secondary structures formed by G-rich DNA or RNA sequences, present in the G-strand of telomeres, resolved by helicases before telomere replication (see Sect. 7.3.1 for details). G-4 stabilizers, like telomestatin isolated from *Streptomyces anulatus*, have been quoted for therapy in preclinical studies. However, the formation of G-quadruplexes motifs is not exclusive of telomeres and off-target action may interfere with replication [230]. Recently another G4 stabilizer has been reported *in vitro* in cultured cells. Combining the natural anthraquinone derivative Emodin (1,3,8-trihydroxy-6-methylanthraquinone) and the small molecule, selective inhibitor of telomerase, BIBR1532 [231], was sufficient to inhibit tumor growth in a murine model [232].

In humans, the TERT enzyme gene comprises approximately 35,000 bp organized into 16 exons and 15 introns, located on the short arm of chromosome 5 (5p15.33) [233]. Important regulatory elements that constitute the structure of its promoter are in a region extending from 330 bp upstream of the ATG translation start codon to the second exon of the gene. This region includes CpG-rich sequences and binding motifs for several transcription factors like SP1 and c-Myc (see Chap. 4). The wide variety of transcription factors that interact with the human *TERT* regulatory region reflects the extensive cellular control over the transcription of this gene [234] (Fig. 7.6). In adult somatic cells, repression of *TERT* is maintained by chromatin remodeling and epigenetic alterations. The majority of cancers acquire replicative immortality through telomerase reactivation, and 10 to 15% of them have mutations in the promoter sequence of the human *TERT* gene [218]. Tumor promoter mutations (TPM) create new binding motifs for transcription factors. The frequent *TERT* somatic short-range mutations C250T (-146C/T) and C228T (-124C/T) substitute a cytosine for a thymidine at positions 250 or 228 of the *TERT* gene promoter-proximal segment, respectively; whereas the negative number terminology in brackets corresponds to the nucleotide position in relation to the transcription start site (see Chap. 4). In both cases, the mutations generate de novo 11-bp sequence (5′-CCCGGAAGGGG-3′) recognized by E26 Transformation-Specific family transcription factor (ETS). In glioblastoma cell lines and orthotopic xenograft the TPM-generated DNA binding site is recognized by GABPA, a component of the multimeric transcription factor GABP that reactivates hTERT [235]. Other transcription factors like ETS1 can reactivate hTERT in melanoma with mild action observed in glioblastoma [236], whereas ETV5 up-regulates hTERT in thyroid cancer cells *in vitro* [237]. These results imply that additional hTERT inhibition/activation might depend on supplementary signaling pathways such as MEK, BRAF, NRAS and NF-κB, chromatin structure and activity (see Chap. 4), and pathological context-dependence [238].

7.6.2 Senescent Cells and Senolytic Drugs

In mammals, telomerase genes' alterations are a risk factor for shortened lifespan or the appearance of age-related diseases. Mice with shortened or elongated telomeres display decreased or increased lifespan, respectively [239]. These experiments

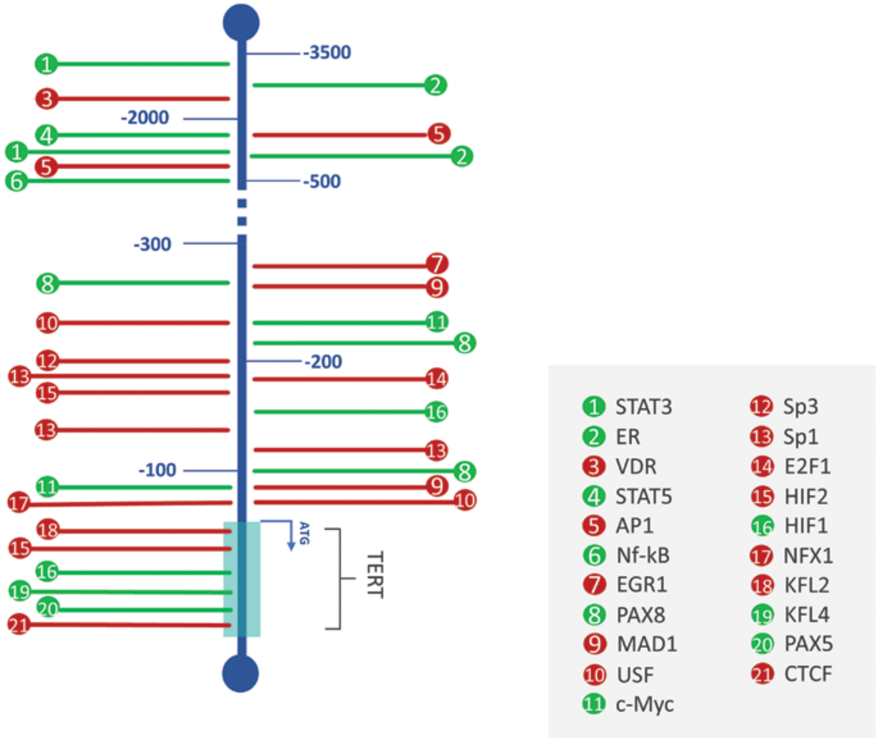


Fig. 7.6 Human *TERT* promoter as a target for cancer. Human *TERT* promoter and binding sites for activators (green) or repressors (red) of transcription upstream the ATG start codon (+1 for transcription) to the second exon of the gene. This region is characterized by a CpG island of about 1138 bp, which extends from position −808 to position +330, with a CG content of 71.3%. In the proximal region of the promoter, from position −721 to position −13, there are about 71 CpG sites, capable of methylation. In hTERT, hypermethylation is related to increased activity of the enzyme telomerase

suggest that lifespan can be manipulated. Indeed, telomere exhaustion explains limited proliferative capacity initially seen in subcultured embryonic fibroblasts [5]. Ectopic expression of telomerase can confer immortality to somatic cells [240], and shortened telomeres are found in healthy aged humans (as well as mice) [3]. Cells with shortened telomeres display characteristics of aging and are called senescent cells. Its accumulation is time-dependent and a trait of aging. Senescent cells are characterized by loss of proliferative potential by regulating different signaling pathways (p38MAPK, JAK2/STAT3, inflammasome, mTOR, ATM/ATR) for the permanent arrest of the cell cycle through hipophosphorylation of retinoblastoma protein and E2F transcription factor. While losing the ability to proliferate is an irreversible condition, cells remain metabolically active and may modulate neighboring cells' phenotype. Secretion of soluble pro-inflammatory mediators (IL-1α/β, IL-6 and TNF-α), reactive species (oxygen and nitrogen) and matrix metalloproteinases denote the Senescence Associated Secretion Phenotype or SASP [241]. Additional characteristics of the SASP phenotype include histone modifications, extensive single- and double-stranded DNA damage, protein carbonylation, lipoperoxidation [196], morphological changes and resistance to apoptosis [242–244].

Senescence is considered a cell fate and a physiological mechanism. It occurs *in vivo* during embryogenesis [245, 246] or acute wound repair. In adult organisms, accumulation of senescent cells is induced by endogenous (e.g., pro-inflammatory or pro-oxidant environment) or exogenous (radiation) insults and function as a tumor-suppressor mechanism, recruiting immune cells to clear particular cells prone for neoplastic transformation. Though transient induction of senescence is beneficial and stimulates tissue regeneration, persistent senescence and the immune system's failure to function properly (exhaustion of stem or progenitor cell compartment) contributes to dysfunctional tissues, pathological aging [247] and can promote cancer resistance. Transgenic models have established causal links between telomere loss, cellular senescence and aging, as a proof of concept. Seminal papers have demonstrated that aging can be reverted by telomerase reactivation [248] or deletion of the *CDKN2A* gene that encodes P16, a marker of senescent cells [249]. The "theory of threshold of senescent cell burden" suggests that tissues can hold a limited number of senescent cells, above which the immune system cannot cope with and eliminate them, leading to the appearance of multimorbidity [250]. The benefits of controlling the accumulation of replicative or prematurely induced senescent cells suggest that lifestyle changes or developing drugs that specifically eliminate senescent cells may represent an attractive target to alleviate the anticipation of age-related disorders or destroy cancer cells. Yet transient induction of senescence is beneficial and stimulates tissue regeneration, persistent senescent cells and the failure of the immune system to properly function (exhaustion of stem or progenitor cell compartment) contributes to dysfunctional tissues, pathological aging [247] and can promote pro-carcinogenic tumor environment (TME) [251]. Ironically, while senescence is initially a tumor-suppressive cellular response, oncogene induced senescence (OIS) and therapy-induced senescence (TIS) become major pro-tumorigenic mechanisms.

Interventions that target essential aging processes such as cellular senescence are being quoted and the use of senolytics is being explored as an anti-aging therapy. For that, senescent cells can substantially contribute to clustered pathological conditions [252] and senolytics may delay, counteract or alleviate multiple age-related diseases. Because senolytics do not target single-molecule receptors, enzymes or biochemical pathways but whole senescent cells, they may be called "panolytic drugs" [253]. Specifically, in cancer, the rationale behind this approach is that after exposure to chemotherapy or radiation new senescent cells appear in the targeted organ, inducing organ-specific premature aging. Though senescent cells can spread, senolytics are usually administered systemically. Metformin, a synthetic analogue of a natural product used in herbal medicine, galegine, is a well-known FDA-approved drug for type II diabetes [254]. Consistent with observations from extended life span in the nematode *Caenorhabditis elegans* [255] recapitulated in mice [256] and the fact that, in humans, metformin is associated with suppression of pro-inflammatory status [257], it has been repurposed for a variety of clinical trials in cancer and aging [254]. Another compound called dasatinib (Bristol-Myers Squibb, New York, NY), a pan-receptor inhibitor of tyrosine kinases, tested in preclinical murine models is now in phase I/II clinical trial as a senolytic drug. Its combination with the natural flavonoid compound quercetin (antioxidant, inhibitor of targeting pro-survival or anti-apoptotic proteins such as members of the BCL-2 family) was tested for diabetic kidney disease (DKD) patients with efficient reduction of senescent cells and SASP mediators [258]. In association with another natural flavonoid, fisetin (antioxidant), dasatinib has alleviated frailty symptoms in IPF, a well-know telomeropathy (see Sect. 7.5) [259].

These results support the idea that, senescence, senescent cells and the SASP phenotype can be manipulated, and disorders can be treated as a group instead of one-at-a-time pathological condition. However, the possible direct or indirect modulation of these new senolytic compounds and their effects on the telomere biology are still open questions.

7.7 Conclusions and Future Directions

The last 35 years have been crucial in demonstrating a key biological role of telomeres in cell homeostasis and aging. The increasing knowledge about telomere biology in model organisms and uncovering the complex and intricate routes that cells use to maintain telomere length at the molecular level hindered many important discoveries. The conserved features that maintain telomere structure highlight the importance of telomeres throughout evolution. Telomere maintenance is mainly provided by the dynamic actions between telomeric proteins, the telomerase complex and the transcription of telomeric lncRNAs (e.g., TERRA), crucial in regulating chromatin and telomere length.

In humans, normal cells keep cell cycle checkpoints and DNA damage sensors correctly operating to avoid DDR and telomere loss. Contrary, the accumulation of damaged DNA and telomere dysfunction emerge as hallmarks of cellular aging. Defects or depletion of telomere components can also be decisive to determine cell fate. Telomere dysfunction, most of the time, results in aging-related syndromic diseases, senescence or death. In this scenario, telomeres represent cumulative experiences of an individual and may represent the interaction between genes, social relations, environment and different types of stress. For that, the topic of telomere biology has opened a completely new field in science. DNA damage accumulation and short telomeres characterize that senescent cells have profound impact on lifespan and healthspan. In the last decade much of the attention has been directed to the study of senescent cells with short telomeres, with the promise to hold a new class of therapeutics targeting directly telomerase or senescent cells. In this regard, future interventions to continue to improve human healthspan and longevity beyond telomerase biology are still to come.

Acknowledgement This work was supported by São Paulo Research Foundation (FAPESP, Fundação de Amparo a Pesquisa do Estado de São Paulo, Brazil, http://www.fapesp.br/), grant 2018/04375-2 to MINC and Federal Agency CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior—Brasil) to FMBT. Morea, E.G.O is a postdoctoral fellow from FAPESP (grant 2019/11496-3); BCDO and MES are doctoral fellows from FAPESP (grants 2019/25985-6, 2020/00316-1); SCP, VLS and VSF are undergraduate fellows from FAPESP (2018/25665-9, 2018/03095-6, 2018/03095-6). MINC is a PQ-CNPq fellow (grant 302433/2019-8). LKG is a postdoctoral fellow from CAPES; VP is a doctoral fellow from CAPES. FMBT is a PQ-CNPq fellow (grant 314375/2018-0). The funders had no role in study design, data collection, analysis, decision to publish or manuscript preparation.

References

1. Muller JH. The remaking of chromosomes. Collect Net. 1938;3:181–98.
2. McClintock B. The behavior in successive nuclear divisions of a chromosome broken at meiosis. Proc Natl Acad Sci U S A. 1939; https://doi.org/10.1073/pnas.25.8.405.
3. Blackburn EH, Gall JG. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. J Mol Biol. 1978; https://doi.org/10.1016/0022-2836(78)90294-2.
4. Szostak JW, Blackburn EH. Cloning yeast telomeres on linear plasmid vectors. Cell. 1982;29:245–55.
5. Hayflick L, Moorhead PS. The serial cultivation of human diploid cell strains. Exp Cell Res. 1961;25:585–621.
6. Watson J. Origin of concatemeric T7DNA. Nature. 1972; https://doi.org/10.1038/10.1038/newbio239197a0.
7. Olovnikov AM. A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. J Theor Biol. 1973; https://doi.org/10.1016/0022-5193(73)90198-7.
8. Hiyama K, et al. Activation of telomerase in human lymphocytes and hematopoietic progenitor cells. J Immunol. 1995;155:3711–5.
9. Harley CB, Futcher AB, Greider CW. Telomeres shorten during ageing of human fibroblasts. Nature. 1990; https://doi.org/10.1038/345458a0.
10. Hastie ND, et al. Telomere reduction in human colorectal carcinoma and with ageing. Nature. 1990;346:866–8.
11. Greider CW, Blackburn EH. Identification of a specific telomere terminal transferase activity in tetrahymena extracts. Cell. 1985;43:405–13.
12. Lundblad V, Szostak JW. A mutant with a defect in telomere elongation leads to senescence in yeast. Cell. 1989; https://doi.org/10.1016/0092-8674(89)90132-3.
13. Armanios M, Blackburn EH. The telomere syndromes. Nat Rev Genet. 2012;13:693–704.
14. Herbert BS, et al. Inhibition of human telomerase in immortal human cells leads to progressive telomere shortening and cell death. Proc Natl Acad Sci U S A. 1999; https://doi.org/10.1073/pnas.96.25.14276.
15. Blackburn EH, Epel ES, Lin J. Human telomere biology: a contributory and interactive factor in aging, disease risks, and protection. Science (80-). 2015;350:1193–8.
16. Meyne J, Ratliff RL, Moyzis RK. Conservation of the human telomere sequence (TTAGGG)(n) among vertebrates. Proc Natl Acad Sci U S A. 1989; https://doi.org/10.1073/pnas.86.18.7049.
17. Blackburn EH, Collins K. Telomerase: an RNP enzyme synthesizes DNA. Cold Spring Harb Perspect Biol. 2011; https://doi.org/10.1101/cshperspect.a003558.
18. Griffith JD, et al. Mammalian telomeres end in a large duplex loop. Cell. 1999; https://doi.org/10.1016/S0092-8674(00)80760-6.

19. Erdel F, et al. Telomere recognition and assembly mechanism of mammalian shelterin. Cell Rep. 2017; https://doi.org/10.1016/j.celrep.2016.12.005.
20. Shay JW, Wright WE. Telomeres and telomerase in normal and cancer stem cells. FEBS Lett. 2010; https://doi.org/10.1016/j.febslet.2010.05.026.
21. De Lange T. T-loops and the origin of telomeres. Nat Rev Mol Cell Biol. 2004; https://doi.org/10.1038/nrm1359.
22. Palm W, De Lange T. How shelterin protects mammalian telomeres. Annu Rev Genet. 2008; https://doi.org/10.1146/annurev.genet.41.110306.130350.
23. Doksani Y, Wu JY, De Lange T, Zhuang X. XSuper-resolution fluorescence imaging of telomeres reveals TRF2-dependent T-loop formation. Cell. 2013; https://doi.org/10.1016/j.cell.2013.09.048.
24. De Lange T. Shelterin-mediated telomere protection. Annu Rev Genet. 2018; https://doi.org/10.1146/annurev-genet-032918-021921.
25. Van Ly D, et al. Telomere loop dynamics in chromosome end protection. Mol Cell. 2018; https://doi.org/10.1016/j.molcel.2018.06.025.
26. Denchi EL, De Lange T. Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. Nature. 2007; https://doi.org/10.1038/nature06065.
27. Xu Y. Chemistry in human telomere biology: structure, function and targeting of telomere DNA/RNA. Chem Soc Rev. 2011; https://doi.org/10.1039/c0cs00134a.
28. Collins K, Mitchell JR. Telomerase in the human organism. Oncogene. 2002; https://doi.org/10.1038/sj/onc/1205083.
29. Tan J, Lan L. The DNA secondary structures at telomeres and genome instability. Cell Biosci. 2020; https://doi.org/10.1186/s13578-020-00409-z.
30. Zaug AJ, Podell ER, Cech TR. Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension in vitro. Proc Natl Acad Sci U S A. 2005; https://doi.org/10.1073/pnas.0504744102.
31. Bhattacharjee A, Wang Y, Diao J, Price CM. Dynamic DNA binding, junction recognition and G4 melting activity underlie the telomeric and genome-wide roles of human CST. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkx878.
32. Martínez P, Blasco MA. Replicating through telomeres: a means to an end. Trends Biochem Sci. 2015; https://doi.org/10.1016/j.tibs.2015.06.003.
33. Eisenberg DTA. An evolutionary review of human telomere biology: the thrifty telomere hypothesis and notes on potential adaptive paternal effects. Am J Hum Biol. 2011; https://doi.org/10.1002/ajhb.21127.
34. Bryan TM, Cech TR. Telomerase and the maintenance of chromosome ends. Curr Opin Cell Biol. 1999; https://doi.org/10.1016/S0955-0674(99)80043-X.
35. Autexier C, Lue NF. The structure and function of telomerase reverse transcriptase. Annu Rev Biochem. 2006; https://doi.org/10.1146/annurev.biochem.75.103004.142412.
36. Nassour J, et al. Autophagic cell death restricts chromosomal instability during replicative crisis. Nature. 2019; https://doi.org/10.1038/s41586-019-0885-0.
37. Shay JW, Wright WE. Telomeres and telomerase: three decades of progress. Nat Rev Genet. 2019; https://doi.org/10.1038/s41576-019-0099-1.
38. Viviescas MA, Cano MIN, Segatto M. Chaperones and their role in telomerase ribonucleoprotein biogenesis and telomere maintenance. Curr Proteomics. 2018; https://doi.org/10.2174/1570164615666180713103133.
39. Rubtsova M, Dontsova O. Human telomerase RNA: telomerase component or more? Biomol Ther. 2020; https://doi.org/10.3390/biom10060873.
40. Nguyen KTTT, Wong JMY. Telomerase biogenesis and activities from the perspective of its direct interacting partners. Cancers. 2020; https://doi.org/10.3390/cancers12061679.
41. Musgrove C, Jansson LI, Stone MD. New perspectives on telomerase RNA structure and function. Wiley Interdisciplinary Reviews: RNA. 2018; https://doi.org/10.1002/wrna.1456.
42. Roake CM, Artandi SE. Regulation of human telomerase in homeostasis and disease. Nat Rev Mol Cell Biol. 2020; https://doi.org/10.1038/s41580-020-0234-z.

43. Wyatt HDM, West SC, Beattie TL. InTERTpreting telomerase structure and function. Nucleic Acids Res. 2010; https://doi.org/10.1093/nar/gkq370.
44. Hukezalie KR, Wong JMY. Structure-function relationship and biogenesis regulation of the human telomerase holoenzyme. FEBS J. 2013; https://doi.org/10.1111/febs.12272.
45. Patrick EM, Slivka JD, Payne B, Comstock MJ, Schmidt JC. Observation of processive telomerase catalysis using high-resolution optical tweezers. Nat Chem Biol. 2020; https://doi.org/10.1038/s41589-020-0478-0.
46. Thompson CAH, Wong JMY. Non-canonical functions of telomerase reverse transcriptase: emerging roles and biological relevance. Curr Top Med Chem. 2020; https://doi.org/10.2174/1568026620666200131125110.
47. Feng J, et al. The RNA component of human telomerase. Science (80-). 1995; https://doi.org/10.1126/science.7544491.
48. Mitchell JR, Wood E, Collins K. A telomerase component is defective in the human disease dyskeratosis congenita. Nature. 1999;402:551–5.
49. Rubtsova M, et al. Protein encoded in human telomerase RNA is involved in cell protective pathways. Nucleic Acids Res. 2018; https://doi.org/10.1093/nar/gky705.
50. Calado RT, et al. A spectrum of severe familial liver disorders associate with telomerase mutations. PLoS One. 2009; https://doi.org/10.1371/journal.pone.0007926.
51. Calado RT, Young NS. Telomere diseases. N Engl J Med. 2009; https://doi.org/10.1056/nejmra0903373.
52. Wu P, Takai H, De Lange T. Telomeric 3′ overhangs derive from resection by Exo1 and apollo and fill-in by POT1b-associated CST. Cell. 2012; https://doi.org/10.1016/j.cell.2012.05.026.
53. Chow TT, Zhao Y, Mak SS, Shay JW, Wright WE. Early and late steps in telomere overhang processing in normal human cells: the position of the final RNA primer drives telomere shortening. Genes Dev. 2012; https://doi.org/10.1101/gad.187211.112.
54. Londoño-Vallejo JA, Der-Sarkissian H, Cazes L, Bacchetti S, Reddel RR. Alternative lengthening of telomeres is characterized by high rates of telomeric exchange. Cancer Res. 2004; https://doi.org/10.1158/0008-5472.CAN-03-4035.
55. Lee M, et al. Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. Nucleic Acids Res. 2014; https://doi.org/10.1093/nar/gkt1117.
56. Episkopou H, et al. Alternative lengthening of telomeres is characterized by reduced compaction of telomeric chromatin. Nucleic Acids Res. 2014; https://doi.org/10.1093/nar/gku114.
57. Udroiu I, Sgura A. Alternative lengthening of telomeres and chromatin status. Genes. 2020; https://doi.org/10.3390/genes11010045.
58. Conomos D, et al. Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. J Cell Biol. 2012;199:893–906.
59. Xu M, et al. Nuclear receptors regulate alternative lengthening of telomeres through a novel noncanonical FANCD2 pathway. Sci Adv. 2019; https://doi.org/10.1126/sciadv.aax6366.
60. Yeager TR, et al. Telomerase-negative immortalized human cells contain a novel type of promyelocytic leukemia (PML) body. Cancer Res. 1999;
61. Fasching CL, Neumann AA, Muntoni A, Yeager TR, Reddel RR. DNA damage induces alternative lengthening of telomeres (ALT)-associated promyelocytic leukemia bodies that preferentially associate with linear telomeric DNA. Cancer Res. 2007; https://doi.org/10.1158/0008-5472.CAN-07-1556.
62. Chung I, Leonhardt H, Rippe K. De novo assembly of a PML nuclear subcompartment occurs through multiple pathways and induces telomere elongation. J Cell Sci. 2011; https://doi.org/10.1242/jcs.084681.
63. Hoang SM, O'Sullivan RJ. Alternative lengthening of telomeres: building bridges to connect chromosome ends. Trends Cancer. 2020; https://doi.org/10.1016/j.trecan.2019.12.009.
64. De Lange T. Shelterin: the protein complex that shapes and safeguards human telomeres. Genes Dev. 2005; https://doi.org/10.1101/gad.1346005.
65. Lim CJ, Zaug AJ, Kim HJ, Cech TR. Reconstitution of human shelterin complexes reveals unexpected stoichiometry and dual pathways to enhance telomerase processivity. Nat Commun. 2017; https://doi.org/10.1038/s41467-017-01313-w.

66. Hanaoka S, Nagadoi A, Nishimura Y. Comparison between TRF2 and TRF1 of their telomeric DNA-bound structures and DNA-binding activities. Protein Sci. 2009; https://doi.org/10.1110/ps.04983705.
67. Rai R, Chen Y, Lei M, Chang S. TRF2-RAP1 is required to protect telomeres from engaging in homologous recombination-mediated deletions and fusions. Nat Commun. 2016; https://doi.org/10.1038/ncomms10881.
68. Diotti R, Loayza D. Shelterin complex and associated factors at human telomeres. Nucleus. 2011; https://doi.org/10.4161/nucl.2.2.15135.
69. Schmutz I, De Lange T. Shelterin. Curr Biol. 2016; https://doi.org/10.1016/j.cub.2016.01.056.
70. De Lange T. How shelterin solves the telomere end-protection problem. Cold Spring Harb Symp Quant Biol. 2010; https://doi.org/10.1101/sqb.2010.75.017.
71. Wang F, et al. The POT1-TPP1 telomere complex is a telomerase processivity factor. Nature. 2007; https://doi.org/10.1038/nature05454.
72. Kibe T, Zimmermann M, de Lange T. TPP1 blocks an ATR-mediated resection mechanism at telomeres. Mol Cell. 2016; https://doi.org/10.1016/j.molcel.2015.12.016.
73. Lei M, Podell ER, Cech TR. Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. Nat Struct Mol Biol. 2004; https://doi.org/10.1038/nsmb867.
74. Chen Y. The structural biology of the shelterin complex. Biol Chem. 2019; https://doi.org/10.1515/hsz-2018-0368.
75. Arango HG. Bioestatística*:* Teórica e computacional. 2009.
76. Janovič T, Stojaspal M, Veverka P, Horáková D, Hofr C. Human telomere repeat binding factor TRF1 replaces TRF2 bound to shelterin core hub TIN2 when TPP1 is absent. J Mol Biol. 2019; https://doi.org/10.1016/j.jmb.2019.05.038.
77. Wellinger RJ. The CST complex and telomere maintenance: the exception becomes the rule. Mol Cell. 2009; https://doi.org/10.1016/j.molcel.2009.10.001.
78. Miyake Y, et al. RPA-like mammalian Ctc1-Stn1-Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. Mol Cell. 2009; https://doi.org/10.1016/j.molcel.2009.08.009.
79. Price CM, et al. Evolution of CST function in telomere maintenance. Cell Cycle. 2010; https://doi.org/10.4161/cc.9.16.12547.
80. Rice C, Skordalakes E. Structure and function of the telomeric CST complex. Comput Struct Biotechnol J. 2016; https://doi.org/10.1016/j.csbj.2016.04.002.
81. Wold MS. Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. Annu Rev Biochem. 1997; https://doi.org/10.1146/annurev.biochem.66.1.61.
82. Chen LY, Redon S, Lingner J. The human CST complex is a terminator of telomerase activity. Nature. 2012; https://doi.org/10.1038/nature11269.
83. Bryan C, Rice C, Harkisheimer M, Schultz DC, Skordalakes E. Structure of the human telomeric Stn1-Ten1 capping complex. PLoS One. 2013; https://doi.org/10.1371/journal.pone.0066756.
84. Chen LY, Majerská J, Lingner J. Molecular basis of telomere syndrome caused by CTC1 mutations. Genes Dev. 2013; https://doi.org/10.1101/gad.222893.113.
85. Simon AJ, et al. Mutations in STN1 cause Coats plus syndrome and are associated with genomic and telomere defects. J Exp Med. 2016; https://doi.org/10.1084/jem.20151618.
86. Amir M, et al. Structural features of nucleoprotein CST/shelterin complex involved in the telomere maintenance and its association with disease mutations. Cell. 2020; https://doi.org/10.3390/cells9020359.
87. Wang F, Stewart J, Price CM. Human CST abundance determines recovery from diverse forms of DNA damage and replication stress. Cell Cycle. 2014; https://doi.org/10.4161/15384101.2014.964100.
88. Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J. Telomeric repeat-containing RNA and RNA surveillance factors at mammalian chromosome ends. Science (80-). 2007; https://doi.org/10.1126/science.1147182.

89. Luke B, et al. The Rat1p 5′ to 3′ exonuclease degrades telomeric repeat-containing RNA and promotes telomere elongation in Saccharomyces cerevisiae. Mol Cell. 2008; https://doi.org/10.1016/j.molcel.2008.10.019.
90. Schoeftner S, Blasco MA. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat Cell Biol. 2008; https://doi.org/10.1038/ncb1685.
91. Bah A, Wischnewski H, Shchepachev V, Azzalin CM. The telomeric transcriptome of Schizosaccharomyces pombe. Nucleic Acids Res. 2012; https://doi.org/10.1093/nar/gkr1153.
92. Cusanelli E, Romero CAP, Chartrand P. Telomeric noncoding RNA TERRA is induced by telomere shortening to nucleate telomerase molecules at short telomeres. Mol Cell. 2013; https://doi.org/10.1016/j.molcel.2013.08.029.
93. Bettin N, Oss Pegorar C, Cusanelli E. The emerging roles of TERRA in telomere maintenance and genome stability. Cell. 2019; https://doi.org/10.3390/cells8030246.
94. Azzalin CM, Lingner J. Telomere functions grounding on TERRA firma. Trends Cell Biol. 2015; https://doi.org/10.1016/j.tcb.2014.08.007.
95. Kwapisz M, Morillon A. Subtelomeric transcription and its regulation. J Mol Biol. 2020; https://doi.org/10.1016/j.jmb.2020.01.026.
96. Nergadze SG, et al. CpG-island promoters drive transcription of human telomeres. RNA. 2009; https://doi.org/10.1261/rna.1748309.
97. Farnung BO, Giulotto E, Azzalin CM. Promoting transcription of chromosome ends. Transcription. 2010; https://doi.org/10.4161/trns.1.3.13191.
98. Pfeiffer V, Lingner J. TERRA promotes telomere shortening through exonuclease 1-mediated resection of chromosome ends. PLoS Genet. 2012; https://doi.org/10.1371/journal.pgen.1002747.
99. Porro A, Feuerhahn S, Reichenbach P, Lingner J. Molecular dissection of telomeric repeat-containing RNA biogenesis unveils the presence of distinct and multiple regulatory pathways. Mol Cell Biol. 2010; https://doi.org/10.1128/mcb.00460-10.
100. Feuerhahn S, Iglesias N, Panza A, Porro A, Lingner J. TERRA biogenesis, turnover and implications for function. FEBS Lett. 2010; https://doi.org/10.1016/j.febslet.2010.07.032.
101. Schoeftner S, Blasco MA. A higher order of telomere regulation: telomere heterochromatin and telomeric RNAs. EMBO J. 2009; https://doi.org/10.1038/emboj.2009.197.
102. Porro A, et al. Functional characterization of the TERRA transcriptome at damaged telomeres. Nat Commun. 2014; https://doi.org/10.1038/ncomms6379.
103. Feretzaki M, Nunes PR, Lingner J. Expression and differential regulation of human TERRA at several chromosome ends. RNA. 2019; https://doi.org/10.1261/rna.072322.119.
104. Redon S, Reichenbach P, Lingner J. The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. Nucleic Acids Res. 2010; https://doi.org/10.1093/nar/gkq296.
105. Yehezkel S, et al. Characterization and rescue of telomeric abnormalities in ICF syndrome type I fibroblasts. Front Oncol. 2013; https://doi.org/10.3389/fonc.2013.00035.
106. Sagie S, et al. Telomeres in ICF syndrome cells are vulnerable to DNA damage due to elevated DNA:RNA hybrids. Nat Commun. 2017; https://doi.org/10.1038/ncomms14015.
107. De Silanes IL, et al. Identification of TERRA locus unveils a telomere protection role through association to nearly all chromosomes. Nat Commun. 2014; https://doi.org/10.1038/ncomms5723.
108. Cusanelli E, Chartrand P. Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity. Front Genet. 2015; https://doi.org/10.3389/fgene.2015.00143.
109. Montero JJ, López De Silanes I, Granã O, Blasco MA. Telomeric RNAs are essential to maintain telomeres. Nat Commun. 2016; https://doi.org/10.1038/ncomms12534.
110. Arora R, Brun CMC, Azzalin CM. TERRA: long noncoding RNA at eukaryotic telomeres. Prog Mol Subcell Biol. 2011;51

111. Deng Z, Norseen J, Wiedmer A, Riethman H, Lieberman PM. TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. Mol Cell. 2009; https://doi.org/10.1016/j.molcel.2009.06.025.
112. Redon S, Zemp I, Lingner J. A three-state model for the regulation of telomerase by TERRA and hnRNPA1. Nucleic Acids Res. 2013; https://doi.org/10.1093/nar/gkt695.
113. Postepska-Igielska A, et al. The chromatin remodelling complex NoRC safeguards genome stability by heterochromatin formation at telomeres and centromeres. EMBO Rep. 2013; https://doi.org/10.1038/embor.2013.87.
114. Montero JJ, et al. TERRA recruitment of polycomb to telomeres is essential for histone trymethylation marks at telomeric heterochromatin. Nat Commun. 2018; https://doi.org/10.1038/s41467-018-03916-3.
115. Chu HP, et al. TERRA RNA antagonizes ATRX and protects telomeres. Cell. 2017; https://doi.org/10.1016/j.cell.2017.06.017.
116. Biffi G, Tannahill D, Balasubramanian S. An intramolecular G-quadruplex structure is required for binding of telomeric repeat-containing RNA to the telomeric protein TRF2. J Am Chem Soc. 2012; https://doi.org/10.1021/ja305734x.
117. Aguado J, d'Adda di Fagagna F, Wolvetang E. Telomere transcription in ageing. Ageing Res Rev. 2020; https://doi.org/10.1016/j.arr.2020.101115.
118. Avogaro L, et al. Live-cell imaging reveals the dynamics and function of single-telomere TERRA molecules in cancer cells. RNA Biol. 2018; https://doi.org/10.1080/15476286.2018.1456300.
119. Collie GW, Haider SM, Neidle S, Parkinson GN. A crystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex. Nucleic Acids Res. 2010; https://doi.org/10.1093/nar/gkq259.
120. Blackburn EH, Greider CW, Szostak JW. Telomeres and telomerase: the path from maize. Tetrahymena and yeast to human cancer and aging. Nat Med. 2006; https://doi.org/10.1038/nm1006-1133.
121. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153:1194–217.
122. Martínez P, Blasco MA. Heart-breaking telomeres. Circ Res. 2018;123:787–802.
123. Takubo K, et al. Telomere shortening with aging in human liver. J Gerontol - Ser A Biol Sci Med Sci. 2000;55:B533–6.
124. Song Z, et al. Lifestyle impacts on the aging-associated expression of biomarkers of DNA damage and telomere dysfunction in human blood. Aging Cell. 2010;9:607–15.
125. Jiang H, et al. Proteins induced by telomere dysfunction and DNA damage represent biomarkers of human aging and disease. Proc Natl Acad Sci U S A. 2008;105:11299–304.
126. Demissie S, et al. Insulin resistance, oxidative stress, hypertension, and leukocyte telomere length in men from the Framingham Heart Study. Aging Cell. 2006;5:325–30.
127. Diez Roux AV, et al. Race/ethnicity and telomere length in the multi-ethnic study of atherosclerosis. Aging Cell. 2009;8:251–7.
128. Hunt SC, et al. Leukocyte telomeres are longer in African Americans than in whites: the National Heart, Lung, and Blood Institute Family Heart Study and the Bogalusa Heart Study. Aging Cell. 2008;7:451–8.
129. Cherkas LF, et al. The association between physical activity in leisure time and leukocyte telomere length. Arch Intern Med. 2008;168:154–8.
130. Lapham K, et al. Automated assay of telomere length measurement and informatics for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics. 2015;200:1061–72.
131. Andrew T, et al. Mapping genetic loci that determine leukocyte telomere length in a large sample of unselected female sibling pairs. Am J Hum Genet. 2006;78:480–6.
132. Levy D, et al. Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. Proc Natl Acad Sci U S A. 2010;107:9293–8.

133. Vasa-Nicotera M, et al. Mapping of a major locus that determines telomere length in humans. Am J Hum Genet. 2005;76:147–51.
134. Mangino M, et al. A regulatory SNP of the BICD1 gene contributes to telomere length variation in humans. Hum Mol Genet. 2008;17:2518–23.
135. Codd V, et al. Common variants near TERC are associated with mean telomere length. Nat Genet. 2010;42:197–9.
136. Takubo K, et al. Telomere lengths are characteristic in each human individual. Exp Gerontol. 2002;37:523–31.
137. Canela A, Vera E, Klatt P, Blasco MA. High-throughput telomere length quantification by FISH and its application to human population studies. Proc Natl Acad Sci U S A. 2007;104:5300–5.
138. Ehrlenbach S, et al. Influences on the reduction of relative telomere length over 10 years in the population-based bruneck study: Introduction of a well-controlled high-throughput assay. Int J Epidemiol. 2009;38:1725–34.
139. Valdes AM, et al. Obesity, cigarette smoking, and telomere length in women. Lancet. 2005;366:662–4.
140. O'Donnell CJ, et al. Leukocyte telomere length and carotid artery intimai medial thickness R the framingham heart study. Arterioscler Thromb Vasc Biol. 2008;28:1165–71.
141. Pavanello S, et al. Shortened telomeres in individuals with abuse in alcohol consumption. Int J Cancer. 2011;129:983–92.
142. Strandberg TE, et al. Association between alcohol consumption in healthy midlife and telomere length in older men. The Helsinki Businessmen Study. 2012. https://doi.org/10.1007/s10654-012-9728-0.
143. Martens DS, et al. Association of parental socioeconomic status and newborn telomere length. JAMA Netw Open. 2020;3:e204057.
144. Kemp BR, Ferraro KF. Are biological consequences of childhood exposures detectable in telomere length decades later? J Gerontol Ser A. 2021; https://doi.org/10.1093/gerona/glaa019.
145. Tucker LA. Physical activity and telomere length in U.S. men and women: an NHANES investigation. Prev Med (Baltim). 2017;100:145–51.
146. Du M, et al. Physical activity, sedentary behavior, and leukocyte telomere length in women. Am J Epidemiol. 2012;175:414–22.
147. Arem H, et al. Leisure time physical activity and mortality: a detailed pooled analysis of the dose-response relationship. JAMA Intern Med. 2015;175:959–67.
148. Starr JM, et al. Oxidative stress, telomere length and biomarkers of physical aging in a cohort aged 79 years from the 1932 Scottish Mental Survey. Mech Ageing Dev. 2008;129:745–51.
149. Von Zglinicki T. Oxidative stress shortens telomeres. Trends Biochem Sci. 2002; https://doi.org/10.1016/S0968-0004(02)02110-2.
150. Ma D, Zhu W, Hu S, Yu X, Yang Y. Association between oxidative stress and telomere length in type 1 and type 2 diabetic patients. J Endocrinol Investig. 2013;36:1032–7.
151. Sampson MJ, Winterbone MS, Hughes JC, Dozio N, Hughes DA. Monocyte telomere shortening and oxidative DNA damage in type 2 diabetes. Diabetes Care. 2006;29:283–9.
152. O'Donovan A, et al. Cumulative inflammatory load is associated with short leukocyte telomere length in the health, aging and body composition study. PLoS One. 2011;6
153. Al-Attas OS, et al. Adiposity and insulin resistance correlate with telomere length in middle-aged Arabs: the influence of circulating adiponectin. Eur J Endocrinol. 2010;163:601–7.
154. Mawanda F, Wallace R. Telomere length in epidemiology: a biomarker of aging, age-related disease, both, or neither? Epidemiol Rev. 2013;35:161–80.
155. Von Zglinicki T, Pilger R, Sitte N. Accumulation of single-strand breaks is the major cause of telomere shortening in human fibroblasts. Free Radic Biol Med. 2000; https://doi.org/10.1016/S0891-5849(99)00207-5.
156. Townsley DM, et al. Danazol treatment for telomere diseases. N Engl J Med. 2016;374:1922–31.

157. Leteurtre F, et al. Accelerated telomere shortening and telomerase activation in Fanconi's anaemia. Br J Haematol. 1999; https://doi.org/10.1046/j.1365-2141.1999.01445.x.
158. Zinsser F. Atrophia cutis reticularis cum pigmentatione, dystrophia unguium et leukoplakia oris. Ikonogr Dermatol (Hyoto). 1910;5:219–23.
159. Costello MJ, Buncke CM. Dyskeratosis congenita. A M A Arch Dermatology. 1956; https://doi.org/10.1001/archderm.1956.01550020023004.
160. Alter BP, Giri N, Savage SA, Rosenberg PS. Cancer in the national cancer institute inherited bone marrow failure syndrome cohort after fifteen years of follow-up. Haematologica. 2018; https://doi.org/10.3324/haematol.2017.178111.
161. Armanios M, et al. Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. Proc Natl Acad Sci U S A. 2005;102:15960–4.
162. Vulliamy TJ, et al. Mutations in dyskeratosis congenita: their impact on telomere length and the diversity of clinical presentation. Blood. 2006;107:2680–5.
163. Aalfs CM, van den Berg H, Barth PG, Hennekam RCM. The Hoyeraal-Hreidarsson syndrome: the fourth case of a separate entity with prenatal growth retardation, progressive pancytopenia and cerebellar hypoplasia. Eur J Pediatr. 1995; https://doi.org/10.1007/BF01957367.
164. Revesz T, Fletcher S, Al-Gazali LI, DeBuse P. Bilateral retinopathy, aplastic anaemia, and central nervous system abnormalities: a new syndrome? J Med Genet. 1992; https://doi.org/10.1136/jmg.29.9.673.
165. Kajtar P, Mehes K. Bilateral Coats retinopathy associated with aplastic anaemia and mild dyskeratotic signs. Am J Med Genet. 1994; https://doi.org/10.1002/ajmg.1320490404.
166. Ehrlich P. Über einen Fall von Anämie mit Bemerkungen über regenerative Veränderungen des Knochenmarks, Charité-Ann. 1888.
167. Arias-Salgado EG, et al. Genetic analyses of aplastic anemia and idiopathic pulmonary fibrosis patients with short telomeres, possible implication of DNA-repair genes. Orphanet J Rare Dis. 2019;14:82.
168. Savage SA, et al. Genetic variation in telomeric repeat binding factors 1 and 2 in aplastic anemia. Exp Hematol. 2006; https://doi.org/10.1016/j.exphem.2006.02.008.
169. Marsh JW, et al. Heterozygous RTEL1 variants in bone marrow failure and myeloid neoplasms. Blood Adv. 2018; https://doi.org/10.1182/bloodadvances.2017008110.
170. Fernández Pérez ER, et al. Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis a population-based study. Chest. 2010;137:129–37.
171. Hodgson U, Laitinen T, Tukiainen P. Nationwide prevalence of sporadic and familial idiopathic pulmonary fibrosis: evidence of founder effect among multiplex families in Finland. Thorax. 2002;57:338–42.
172. Kim HJ, Perlman D, Tomic R. Natural history of idiopathic pulmonary fibrosis. Respir Med. 2015; https://doi.org/10.1016/j.rmed.2015.02.002.
173. Armanios MY, et al. Telomerase mutations in families with idiopathic pulmonary fibrosis. N Engl J Med. 2007;356:1317–26.
174. Tsakiri KD, et al. Adult-onset pulmonary fibrosis caused by mutations in telomerase. Proc Natl Acad Sci U S A. 2007;104:7552–7.
175. Stuart BD, et al. Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. Nat Genet. 2015; https://doi.org/10.1038/ng.3278.
176. Fingerlin TE, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet. 2013; https://doi.org/10.1038/ng.2609.
177. Alder JK, et al. Exome sequencing identifi es mutant TINF2 in a family with pulmonary fibrosis. Chest. 2015;147:1361–8.
178. Alder JK, et al. Short telomeres are a risk factor for idiopathic pulmonary fibrosis. Proc Natl Acad Sci U S A. 2008; https://doi.org/10.1073/pnas.0804280105.
179. Parry EM, Alder JK, Qi X, Chen JJL, Armanios M. Syndrome complex of bone marrow failure and pulmonary fibrosis predicts germline defects in telomerase. Blood. 2011; https://doi.org/10.1182/blood-2010-11-322149.

180. De Sandre-Giovannoli A, et al. Lamin A truncation in Hutchinson-Gilford progeria. Science (80-). 2003;300:2055.
181. Eriksson M, et al. Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. Nature. 2003;423:293–8.
182. Benson EK, Lee SW, Aaronson SA. Role of progerin-induced telomere dysfunction in HGPS premature cellular senescence. J Cell Sci. 2010; https://doi.org/10.1242/jcs.067306.
183. Decker ML, Chavez E, Vulto I, Lansdorp PM. Telomere length in Hutchinson-Gilford Progeria Syndrome. Mech Ageing Dev. 2009; https://doi.org/10.1016/j.mad.2009.03.001.
184. Crabbe L, Cesare AJ, Kasuboski JM, Fitzpatrick JAJ, Karlseder J. Human telomeres are tethered to the nuclear envelope during postmitotic nuclear assembly. Cell Rep. 2012; https://doi.org/10.1016/j.celrep.2012.11.019.
185. Chojnowski A, et al. Progerin reduces LAP2α-telomere association in hutchinson-gilford progeria. elife. 2015; https://doi.org/10.7554/eLife.07759.
186. Wood AM, et al. TRF2 and lamin A/C interact to facilitate the functional organization of chromosome ends. Nat Commun. 2014; https://doi.org/10.1038/ncomms6467.
187. Wood LD, et al. Characterization of ataxia telangiectasia fibroblasts with extended life-span through telomerase expression. Oncogene. 2001; https://doi.org/10.1038/sj.onc.1204072.
188. Crabbe L, Verdun RE, Haggblom CI, Karlseder J. Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. Science (80-). 2004; https://doi.org/10.1126/science.1103619.
189. Collopy LC, et al. Triallelic and epigenetic-like inheritance in human disorders of telomerase. Blood. 2015; https://doi.org/10.1182/blood-2015-03-633388.
190. Brouilette S, Singh RK, Thompson JR, Goodall AH, Samani NJ. White cell telomere length and risk of premature myocardial infarction. Arterioscler Thromb Vasc Biol. 2003; https://doi.org/10.1161/01.ATV.0000067426.96344.32.
191. Codd V, et al. Identification of seven loci affecting mean telomere length and their association with disease. Nat Genet. 2013; https://doi.org/10.1038/ng.2528.
192. Révész D, Milaneschi Y, Verhoeven JE, Penninx BWJH. Telomere length as a marker of cellular aging is associated with prevalence and progression of metabolic syndrome. J Clin Endocrinol Metab. 2014; https://doi.org/10.1210/jc.2014-1851.
193. Révész D, et al. Associations between cellular aging markers and metabolic syndrome: findings from the cardia study. J Clin Endocrinol Metab. 2018; https://doi.org/10.1210/jc.2017-01625.
194. Sanders JL, et al. Leukocyte telomere length is associated with noninvasively measured age-related disease: the cardiovascular health study. J Gerontol Ser A Biol Sci Med Sci. 2012;67(A):409–16.
195. Friis-Ottessen M, et al. Telomere shortening correlates to dysplasia but not to DNA aneuploidy in longstanding ulcerative colitis. BMC Gastroenterol. 2014; https://doi.org/10.1186/1471-230X-14-8.
196. Grun LK, et al. TRF1 as a major contributor for telomeres' shortening in the context of obesity. Free Radic Biol Med. 2018;129
197. Laimer M, et al. Telomere length increase after weight loss induced by bariatric surgery: results from a 10 year prospective study. Int J Obes. 2016; https://doi.org/10.1038/ijo.2015.238.
198. Formichi C, et al. Weight loss associated with bariatric surgery does not restore short telomere length of severe obese patients after 1 year. Obes Surg. 2014;24:2089–93.
199. Bonfigli AR, et al. Leukocyte telomere length and mortality risk in patients with type 2 diabetes. Oncotarget. 2016; https://doi.org/10.18632/oncotarget.10615.
200. Aulinas A, et al. Dyslipidemia and chronic inflammation markers are correlated with Telomere Length shortening in Cushing's syndrome. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0120185.
201. Strazhesko I, et al. Association of insulin resistance, arterial stiffness and telomere length in adults free of cardiovascular diseases. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0136676.

202. Albrecht E, et al. Telomere length in circulating leukocytes is associated with lung function and disease. Eur Respir J. 2014;43:983–92.
203. Córdoba-Lanús E, et al. Telomere shortening and accelerated aging in COPD: findings from the BODE cohort. Respir Res. 2017; https://doi.org/10.1186/s12931-017-0547-4.
204. Kyoh S, et al. Are leukocytes in asthmatic patients aging faster? A study of telomere length and disease severity. J Allergy Clin Immunol. 2013;132
205. Belsky DW, et al. Is chronic asthma associated with shorter leukocyte telomere length at midlife? Am J Respir Crit Care Med. 2014;190:384–91.
206. Scarabino D, et al. Leukocyte telomere shortening in Huntington's disease. J Neurol Sci. 2019; https://doi.org/10.1016/j.jns.2018.10.024.
207. Jing ZG, et al. A percentage analysis of the telomere length in Parkinson's disease patients. J Gerontol Ser A Biol Sci Med Sci. 2008;63:467–73.
208. Silva PNO, et al. Promoter methylation analysis of SIRT3, SMARCA5, HTERT and CDH1 genes in aging and Alzheimer's disease. J Alzheimers Dis. 2008;13:173–6.
209. Zhan Y, et al. Telomere length shortening and Alzheimer disease-A mendelian randomization study. JAMA Neurol. 2015; https://doi.org/10.1001/jamaneurol.2015.1513.
210. Barbé-Tuana FM, et al. Shortened telomere length in bipolar disorder: a comparison of the early and late stages of disease. Rev Bras Psiquiatr. 2016;38:281–6.
211. Vasconcelos-Moreno MPMP, et al. Telomere length, oxidative stress, inflammation and BDNF levels in siblings of patients with bipolar disorder: implications for accelerated cellular aging. Int J Neuropsychopharmacol. 2017;20:445–54.
212. Czepielewski LSLS, et al. Telomere length and CCL11 levels are associated with gray matter volume and episodic memory performance in schizophrenia: evidence of pathological accelerated aging. Schizophr Bull. 2017;44:1–10.
213. Czepielewski LSLS, et al. Telomere length in subjects with schizophrenia, their unaffected siblings and healthy controls: evidence of accelerated aging. Schizophr Res. 2016;174:39–42.
214. Buxton JL, et al. Childhood obesity is associated with shorter leukocyte telomere length. J Clin Endocrinol Metab. 2011; https://doi.org/10.1210/jc.2010-2924.
215. Li Z, He Y, Wang D, Tang J, Chen X. Association between childhood trauma and accelerated telomere erosion in adulthood: a meta-analytic study. J Psychiatr Res. 2017; https://doi.org/10.1016/j.jpsychires.2017.06.002.
216. Mayer SE, et al. Cumulative lifetime stress exposure and leukocyte telomere length attrition: the unique role of stressor duration and exposure timing. Psychoneuroendocrinology. 2019; https://doi.org/10.1016/j.psyneuen.2019.03.002.
217. Clemente DBP, et al. Prenatal and childhood traffic-related air pollution exposure and telomere length in european children: The HELIX project. Environ Health Perspect. 2019; https://doi.org/10.1289/EHP4148.
218. Chiba K, et al. Mutations in the promoter of the telomerase gene TERT contribute to tumorigenesis by a two-step mechanism. Science (80-). 2017; https://doi.org/10.1126/science.aao0535.
219. Kim NW, et al. Specific association of human telomerase activity with immortal cells and cancer. Science. 1994;266:2011–5.
220. Barthel FP, et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. Nat Genet. 2017; https://doi.org/10.1038/ng.3781.
221. Nguyen THD, et al. Cryo-EM structure of substrate-bound human telomerase holoenzyme. Nature. 2018; https://doi.org/10.1038/s41586-018-0062-x.
222. Vonderheide RH. Telomerase as a universal tumor-associated antigen for cancer immunotherapy. Oncogene. 2002; https://doi.org/10.1038/sj/onc/1205074.
223. Lilleby W, et al. Phase I/IIa clinical trial of a novel hTERT peptide vaccine in men with metastatic hormone-naive prostate cancer. Cancer Immunol Immunother. 2017; https://doi.org/10.1007/s00262-017-1994-y.
224. Middleton G, et al. Gemcitabine and capecitabine with or without telomerase peptide vaccine GV1001 in patients with locally advanced or metastatic pancreatic cancer (TeloVac):

an open-label, randomised, phase 3 trial. Lancet Oncol. 2014; https://doi.org/10.1016/S1470-2045(14)70236-0.

225. Seidel JA, Otsuka A, Kabashima K. Anti-PD-1 and anti-CTLA-4 therapies in cancer: mechanisms of action, efficacy, and limitations. Front Oncol. 2018; https://doi.org/10.3389/fonc.2018.00086.
226. Nemunaitis J, et al. A phase i study of telomerase-specific replication competent oncolytic adenovirus (telomelysin) for various solid tumors. Mol Ther. 2010; https://doi.org/10.1038/mt.2009.262.
227. Mender I, Gryaznov S, Dikmen ZG, Wright WE, Shay JW. Induction of telomere dysfunction mediated by the telomerase substrate precursor 6-thio-2′-deoxyguanosine. Cancer Discov. 2015; https://doi.org/10.1158/2159-8290.CD-14-0609.
228. Chiappori AA, et al. A randomized phase II study of the telomerase inhibitor imetelstat as maintenance therapy for advanced non-small-cell lung cancer. Ann Oncol. 2015; https://doi.org/10.1093/annonc/mdu550.
229. Baerlocher GM, Burington B, Snyder DS. Telomerase inhibitor imetelstat in essential thrombocythemia and myelofibrosis. N Engl J Med. 2015;373:2579–81.
230. Sinclair D, Fillman SG, Webster MJ, Weickert CS. Dysregulation of glucocorticoid receptor co-factors FKBP5, BAG1 and PTGES3 in prefrontal cortex in psychotic illness. Sci Rep. 2013;3:3539.
231. Bryan C, et al. Structural basis of telomerase inhibition by the highly specific BIBR1532. Structure. 2015; https://doi.org/10.1016/j.str.2015.08.006.
232. Liu C, Zhou H, Sheng XB, Liu XH, Chen FH. Design, synthesis and SARs of novel telomerase inhibitors based on BIBR1532. Bioorg Chem. 2020; https://doi.org/10.1016/j.bioorg.2020.104077.
233. Bryce LA, Morrison N, Hoare SF, Muir S, Keith WN. Mapping of the gene for the human telomerase reverse transcriptase, hTERT, to chromosome 5p15.33 by fluorescence in situ hybridization. Neoplasia. 2000; https://doi.org/10.1038/sj.neo.7900092.
234. Ramlee MK, Wang J, Toh WX, Li S. Transcription regulation of the human telomerase reverse transcriptase (hTERT) gene. Genes. 2016; https://doi.org/10.3390/genes7080050.
235. Mancini A, et al. Disruption of the β1L isoform of GABP reverses glioblastoma replicative immortality in a TERT promoter mutation-dependent manner. Cancer Cell. 2018; https://doi.org/10.1016/j.ccell.2018.08.003.
236. Vallarelli AF, et al. TERT promoter mutations in melanoma render TERT expression dependent on MAPK pathway activation. Oncotarget. 2016; https://doi.org/10.18632/oncotarget.10634.
237. Bullock M, et al. ETS factor ETV5 activates the mutant telomerase reverse transcriptase promoter in thyroid cancer. Thyroid. 2019; https://doi.org/10.1089/thy.2018.0314.
238. Li Y, Cheng HS, Chng WJ, Tergaonkar V, Cleaver JE. Activation of mutant TERT promoter by RAS-ERK signaling is a key step in malignant progression of BRAF-mutant human melanomas. Proc Natl Acad Sci U S A. 2016; https://doi.org/10.1073/pnas.1611106113.
239. Blasco MA, et al. Telomere shortening and tumor formation by mouse cells lacking telomerase RNA. Cell. 1997; https://doi.org/10.1016/S0092-8674(01)80006-4.
240. Bodnar AG, et al. Extension of life-span by introduction of telomerase into normal human cells. Science (80-). 1998; https://doi.org/10.1126/science.279.5349.349.
241. Coppé JP, et al. Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. PLoS Biol. 2008; https://doi.org/10.1371/journal.pbio.0060301.
242. Wang E. Regulation of apoptosis resistance and ontogeny of age-dependent diseases. Exp Gerontol. 1997; https://doi.org/10.1016/S0531-5565(96)00156-8.
243. Campisi J. Aging, cellular senescence, and cancer. Annu Rev Physiol. 2013; https://doi.org/10.1146/annurev-physiol-030212-183653.
244. Childs BG, Baker DJ, Kirkland JL, Campisi J, Deursen JM. Senescence and apoptosis: dueling or complementary cell fates? EMBO Rep. 2014; https://doi.org/10.15252/embr.201439245.

245. Jeon OH, et al. Local clearance of senescent cells attenuates the development of post-traumatic osteoarthritis and creates a pro-regenerative environment. Nat Med. 2017;23:775–81.
246. Storer M, et al. XSenescence is a developmental mechanism that contributes to embryonic growth and patterning. Cell. 2013; https://doi.org/10.1016/j.cell.2013.10.041.
247. Sharpless NE, Sherr CJ. Forging a signature of in vivo senescence. Nat Rev Cancer. 2015;15:397–408.
248. Jaskelioff M, et al. Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. Nature. 2011; https://doi.org/10.1038/nature09603.
249. Baker DJ, et al. Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. Nature. 2011;479:232–6.
250. Prata LGPL, Ovsyannikova IG, Tchkonia T, Kirkland JL. Senescent cell clearance by the immune system: emerging therapeutic opportunities. Semin Immunol. 2018; https://doi.org/10.1016/j.smim.2019.04.003.
251. Coppé J-P, Desprez P-Y, Krtolica A, Campisi J. The senescence-associated secretory phenotype: the dark side of tumor suppression. Annu Rev Pathol. 2010;5:99–118.
252. Jeck WR, Siebold AP, Sharpless NE. Review: a meta-analysis of GWAS and age-associated diseases. Aging Cell. 2012;11:727–31.
253. Wissler Gerdes EO, Zhu Y, Tchkonia T, Kirkland JL. Discovery, development, and future application of senolytics: theories and predictions. FEBS J. 2020; https://doi.org/10.1111/febs.15264.
254. Heckman-Stoddard BM, DeCensi A, Sahasrabuddhe VV, Ford LG. Repurposing metformin for the prevention of cancer and cancer recurrence. Diabetologia. 2017; https://doi.org/10.1007/s00125-017-4372-6.
255. Cabreiro F, et al. Metformin retards aging in C elegans by altering microbial folate and methionine metabolism. Cell. 2013; https://doi.org/10.1016/j.cell.2013.02.035.
256. Martin-Montalvo A, et al. Metformin improves healthspan and lifespan in mice. Nat Commun. 2013; https://doi.org/10.1038/ncomms3192.
257. Vasamsetti SB, et al. Metformin inhibits monocyte- To-macrophage differentiation via AMPK-mediated inhibition of STAT3 activation: potential role in atherosclerosis. Diabetes. 2015; https://doi.org/10.2337/db14-1225.
258. Hic LJ, et al. Senolytics decrease senescent cells in humans: preliminary report from a clinical trial of Dasatinib plus Quercetin in individuals with diabetic kidney disease. EBioMedicine. 2019;
259. Justice JN, et al. Senolytics in idiopathic pulmonary fibrosis: results from a first-in-human, open-label, pilot study. EBioMedicine. 2019; https://doi.org/10.1016/j.ebiom.2018.12.052.

Chapter 8
To Build or To Break: The Dual Impact of Interspersed Transposable Elements in Cancer

Daniel Andrade Moreira, Cristóvão Antunes de Lanna, Jéssica Gonçalves Vieira da Cruz, and Mariana Boroni

8.1 Transposable Elements Diversity

First discovered by Barbara McClintock in her seminal work on maize [1, 2], interspersed transposable elements (TEs) are genetic elements that can move within a host genome and often duplicate themselves in the process. Although these elements were once described as "junk DNA" or "selfish elements", they are now being recognized as evolutionary toolkits. In this chapter, we have performed a thorough literature revision on how the TEs have shaped the structure, function, and evolution of genomes, with a focus on the human genome. We also dig in the impact of TEs on cancer onset and progression.

TEs are virtually present in all eukaryotic genomes and comprehend a large fraction of the total DNA content of an organism. The composition of TEs in the genomes varies among species, and the increasing availability of whole-genome sequences from diverse organisms, accompanied by the development of genomic techniques targeting transposable elements, has fueled the discovery of large numbers of elements in a wide range of organisms. In mammalian genomes, TEs make up between one to two-thirds of total DNA content [3, 4]. It is important to note that,

D. A. Moreira
Bioinformatics and Computational Biology Lab, Division of Experimental and Translational Research, Brazilian National Cancer Institute, Rio de Janeiro, RJ, Brazil

Functional Genomics and Bioinformatics Lab, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, RJ, Brazil
e-mail: daniel.andrade@fiocruz.br

C. A. de Lanna · J. G. V. da Cruz · M. Boroni (✉)
Bioinformatics and Computational Biology Lab, Division of Experimental and Translational Research, Brazilian National Cancer Institute, Rio de Janeiro, RJ, Brazil
e-mail: cristovao.lanna@inca.gov.br; jessica.granada@inca.gov.br; mariana.boroni@inca.gov.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_8

currently, TE content is yet underestimated due to sequence degradation and methodological limitations [5]. Although the composition of major classes is similar among mammals, even closely related species may have different families of TEs in their genomes and these differences might have played an important role in the molecular differentiation of mammals [6]. For example, one study comparing TE insertions among 29 families of the Mammalia class showed high rates of speciation associated with the high density of TE insertions in the genomes [7].

Regarding human genomes, the first draft has revealed that TEs account for approximately 45% of the total DNA content [8] and, recently, with an alternative approach for repetitive elements detection, de Koning and colleagues (2011) predicted additional repetitive sequences, suggesting that up to 69% of the human genome is repetitive, most of them derived from TEs [9].

Historically, these elements are classified into two major classes, based on their mechanism of transposition [10]: Class I, also known as retrotransposons, and Class II, known as DNA transposons. Class I elements move within the host genome through a copy-and-paste mechanism requiring an RNA intermediate for transposition and reverse transcriptase activity. On the other hand, Class II TEs code for a transposase and do not require an RNA intermediate. All their transposing cycle occurs at the DNA level, usually via a cut-and-paste mechanism involving the excision and reinsertion of the DNA sequence. The detailed description of transposition's mechanism for each of those elements has been extensively reviewed elsewhere [11] and is beyond the scope of this chapter. The classification of transposable elements has long been discussed [12–15] and the used nomenclature depends on the machinery of transposition, phylogeny, and structure.

DNA transposons represent a relatively small fraction (3–4%) of the repetitive content in the human genome (Table 8.1). Yet, they are a very diverse group. According to the Repbase repository [16], the most commonly used database of

Table 8.1 Classification of transposable elements and the proportion of each class in the human genome. TE content for the *Homo sapiens* reference genome (version hg38) were obtained from the RepeatMasker website (http://www.repeatmasker.org/species/hg.html)

	Class	Group	Clade/Superfamily	Proportion of TEs
Class I: Retrotransposon	LTRs		*ERV1, ERV2, ERV3, Gypsy, DIRS*	9.3%
	Non-LTRs	LINEs	*CR1, Crack, L1, L2, R4, RTE, RTEX, Tx1, Vingi, Penelope*	21.8%
		SINEs	*SINE1/7SL, SINE2/tRNA, SINE3/5S*	13.4%
		Composite	*SVA*	0.1%
Class II: DNA Transposon		DDE	*Ginger1, Harbinger, hAT, Kolobok, Mariner/Tc1, Merlin, MuDR, P, piggyBac, Transib*	3.7%
		YR	*Crypton*	
		HUH	*Helitron*	

DDD/E transposase, *YR* tyrosine recombinase, *HUH* endonuclease repetitive DNA elements, they are currently classified into 23 superfamilies, of which 12 can be found in the human genome [17]. The eukaryotic TEs can be grouped into four major groups depending on the transposition machinery and the composition of protein domains encoded by these elements. The dominant group of DNA transposons encodes a DDD/E transposase and it is able to mobilize itself through the classic cut-and-paste mechanism; the second group includes the TE named *Polinton*, also called *Maverick*, which encodes the same DDD/E transposase, in addition to a DNA polymerase B, which is used to its self-replication; the third group includes DNA transposons that encode a tyrosine recombinase, known as *Crypton*, and is proposed to mobilize itself via a circular DNA intermediate; the fourth group is characterized by DNA transposons that encode a HUH nuclease, known as *Helitrons* (rolling-circle transposons) (reviewed in [15]). All transposons that encode DDD/E transposase are characterized by terminal inverted repeats (TIRs), whereas *Cryptons* and *Helitrons* display other short terminal repeats than TIRs. There are also non-autonomous DNA transposons, derived from the DDE/E transposase group, that only contain short terminal fragments and lack coding sequences, such as the miniature inverted-repeat transposable elements (MITEs). Due to the lack of encoded enzymes, they must rely on autonomous DNA transposons for their transposition.

DNA transposons were believed to no longer being able to move in mammals' genomes, including humans, since they were inactivated by mutations, though they were active during early primate evolution until ~37 million years ago, and have impacted human biology with the domestication of some of these elements (reviewed in [4]). An exception, however, has been found in bats from the genus *Myotis*, in which the evidence of recent (~8–30 Ma) DDD/E-derived DNA transposition activity was first documented [18, 19]. Some of those ancient transposons that were once active in the human genome in the past were tamed and contributed to human genome evolution by providing protein-coding sequences. For example, the human diversity of antigen receptors responsible for adaptive immunity relies on the somatic V(D)J recombination, which is dependent on the recombinase encoded by the recombination-activating genes, *RAG1* and *RAG2*. Studies have shown that both *RAG1* and *RAG2* evolved from an ancient DNA transposon (Transib superfamily) [20–22]. Two other examples of human genes derived from DNA transposons are *PGBD5* and *THAP9*. Although domesticated, the proteins encoded by these genes still retain the transposase activity. PiggyBac transposable element-derived protein 5 (*PGBD5*) was domesticated approximately 500 million years ago, in the common ancestor of cephalochordates and vertebrates, and is found to be highly expressed in neurons, suggesting that they may have played a role in the vertebrate nervous system development [23, 24]. The THAP domain–containing 9 (*THAP9*) is a gene encoding an active DNA transposase derived from *P*-element transposase that was found to mobilize transposons in *Drosophila* and human cells. *THAP9*-related genes are widely distributed in eukaryotic genomes with yet unknown function [25].

On the other hand, retrotransposons can be subdivided into two large groups: long-terminal repeat (LTR) elements and non-LTR elements. They represent the vast majority of TE-derived sequences in most genomes, accounting for approximately 45% of total DNA content in the human genome (Table 8.1). LTR retrotransposons are subdivided into nine superfamilies (*Copia*, *Gypsy*, *BEL, DIRS,* and five endogenous retroviruses—*ERV1*, *ERV2*, *ERV3*, *ERV4*, and *Lentivirus*). The most distinctive LTR retrotransposons are the members of the *DIRS* group (or tyrosine recombinase-encoding retrotransposons): they are located as a branch inside of LTR retrotransposons in the reverse transcriptase phylogeny, although they do not have the LTR sequences. Instead, they have either split repeats (SRs) or inverted terminal repeats (ITRs) [26]. Though the origin of this group has not been appropriately determined, there is evidence that they were generated via recombination between a Crypton-like TE and an LTR retrotransposon [15], and only recently their mechanism of transposition was elucidated [27].

In humans, there is evidence of the presence of five superfamilies of LTR elements (*ERV1*, *ERV2*, *ERV3*, *Gypsy*, and *DIRS*), with a predominance of human ERVs (HERVs), and some traces of domesticated *Gypsy* LTR retrotransposon [17]. HERVS are characterized by long sequences (up to 10,000 bp) and account for ~8–10% of the human genome (Table 8.1) [28]. Like DNA transposons, almost all HERVs lost their ability to retrotranspose within the genome [29], with the exception of *HERVK* that exhibits polymorphic insertions in the human population and seems to be still active [30]. Although the majority of LTR elements are no longer active, there is evidence of the exaptation of LTR-derived ORFs encoding essential proteins for mammalian development. For instance, proteins called *syncytins*, which are derived from the envelope glycoprotein-encoding (*env*) genes of endogenous retroviruses, contribute to the formation of the placental cell-cell fused layer called the syncytiotrophoblast, at the fetal-maternal interface [31]. However, the new functions of these elements are more frequently found co-opted as regulatory sequences controlling host gene expression. For example, upon IFN-γ exposure, macrophages show induction of histone H3 Lys^{27} acetylation on STAT1 binding sites regions that were derived from *MER41A* (a member of the *ERV1* family) LTRs. This region acts as enhancers of IFN-stimulated genes, which are important for innate immune responses [32].

The non-LTR elements encode apurinic-like endonuclease (APE) and/or restriction-like endonuclease (RLE), and usually have poly(A) or simple repeats at their 3′-terminus. The replication process of non-LTR retrotransposons is known as target-primed reverse transcription and it is promoted by annealing poly(A) tails to T-rich sites in the genome [11]. These elements are further divided into two major classes, the long interspersed nuclear elements (LINEs) and the short interspersed nuclear elements (SINEs), and one composite retrotransposon *SINE-R/VNTR/Alu* (*SVA*s) group [17]. This adopted classification of non-LTR retrotransposons is usually phylogeny-based and subdivides each major class in clades (LINEs: 32 clades, including *Penelope*-like elements, that can either encode or not the GIY-YIG nuclease; SINEs: 5 clades, including the *Alu* elements, which belong to the *SINE1/7SL* clade and are the most abundant transposable element in the human genome; and *SVA*s) [15]. The human genome contains traces of 10 clades of LINEs (*L1*, *CR1*, *L2*, *Crack*, *RTE*, *RTEX*, *R4*, *Vingi*, *Tx1*, and *Penelope*), three types of SINEs (*SINE1/7SL*, *SINE2/tRNA*, and *SINE3/5S*), and one composite retrotransposon *SVA* (Table 8.1)

[17]. Together, these elements account for ~73% of human TEs, and over 1 billion bp, one-third of the whole human genome [33].

Only three groups of non-LTR TEs are still active in the human genome: *L1* (long interspersed element-1 (LINE-1)), *Alu,* and *SVA*. Their recent activity has seemingly contributed to our differentiation as a species considering the enrichment of gene variants co-opted in hominid genomes. In a comparison of retrotransposon insertions differentially present in the genomes of Anatomically Modern Humans, Neanderthals, Denisovans, and Chimpanzees, the authors found that the expression of genes containing the most recent non-LTR insertions (specific of Modern Humans) was enriched in the brain, where they are related to neuron maturation and migration [34]. It is currently estimated that *Alu*, *L1*, and *SVA* germline transposition events occur at a rate of 1:20 to 1:200 births [35].

The canonical, full-length *L1* element is ~6 kb long and consists of two open reading frames (ORFs) flanked by 5′ and 3′ untranslated regions, with an internal RNA polymerase II promoter and a polyadenylation signal, respectively. ORF1 encodes an RNA-binding protein and ORF2 encodes a protein with endonuclease and reverse-transcriptase activities [36]. *L1* is the only autonomously active family in humans, constituting approximately 17–20% of the human genome, including more than 500,000 copies [37]. Interestingly, ~99.9% of these copies are fixed in the genome due to the accumulation of various 5′ truncated forms that are associated with premature reverse transcription termination, internal rearrangements, and mutations, making them no longer active. This resulted in only a few copies potentially active in the genome [36]. However, the abundance of active *L1*s in the human population remains largely unexplored. One study using a cell retrotransposition assay reported 68 full-length *L1*s (*L1HS* subfamily) differentially present among the population that are absent from the human genome reference sequence. They also showed that the majority of *L1*s were highly active in modern-day human genomes [38]. Nonetheless, recent results reported that *L1HS* transcription is predominantly inactive in somatic human cells, but a small number of copies can escape silencing and be transcriptionally activated in somatic cells regulated by individual-, locus-, and cell-type-specific determinants [39].

Compared to *L1*, *Alu* elements, a short interspersed element, constitute a smaller portion of the human genome (~11%), yet totalizing more than one million copies [40], making this TE the most frequent in the human genome in terms of copy number. *Alus* are primate-specific repeats and the typical full-length structure is ~300 bp-long. Their structure is dimeric and formed by the fusion of two monomers derived from the *7SL RNA* gene [41, 42]. The 5′ region contains an internal RNA polymerase III promoter and a polyadenylation signal in the terminal portion. As these elements do not have coding capacity, they are considered non-autonomous TEs, and their retrotransposition relies on the *L1* molecular machinery [37]. Members of *AluY* and *AluS* were shown to be transposition-competent, showing polymorphic distributions in the human population [43, 44]. A recent study showed that *Alu* sequences exaptation gave rise to active regulators of gene transcription. Interestingly, *Alu* insertions located in genic regions (3′ UTR and proximal promoter regions) were commonly associated with increased gene expression [45].

Much less abundant than *L1* and *Alu*, *SVA* composite elements (hominid-specific element) only make up ~0.2% (~2700 copies) of the human genome [40, 46]. The full-length *SVA* element is ~2 kb long and is composed of a $(CCCTCT)_n$ hexamer repeat region; an *Alu*-like region; a variable number of tandem repeats (VNTR); a *SINE-R* domain derived from the 3′ end of a human endogenous retrovirus K10 (*HERV-K10*); and a polyadenylation signal in the final portion. They apparently lack internal RNA polymerase promoter and, like *Alu* elements, *SVA*s are non-autonomous TEs, that also rely on the *L1* retrotransposition machinery [47].

Notwithstanding the TEs' contribution to novel and beneficial host functions, such as enhancing the diversity of coding and regulatory sequences in the genomes, or enabling structural variations, contributing substantially to genome evolution, the activity of some TEs are also considered a threat to genomic integrity. Somatic transposition events might result in deleterious mutations and be responsible for diseases that will be further discussed in this chapter. Therefore, the host genome has evolved several strategies to control transposition. In this regard, TEs have been highlighted as a "double-edged sword", whereby host genomes must keep and co-opt them to provide functional benefits while, at the same time, hamper deleterious events that disrupt gene function and contribute to genomic instability in a multidimensional scale [48]. We will see how the host genome defends itself from TEs mobilization, and the impacts and consequences of TEs transposition.

8.2 Transposable Elements' Regulation

In this section we will see the host genome strategies to repress TEs activity, trying to avoid deleterious effects that can result from their transposition. The activity of TEs in the human genome is regulated in multiple layers, from transcriptional repression by gene silencing through co-transcriptional regulation, to a post-transcriptional control regulated by RNA editing and degradation. Transcriptional repression is a major mechanism of defense against retrotransposons and is a dynamic process depending on the embryonic development stage, TEs class, and cell type [49]. These mechanisms that repress TE activity are especially important in germline cells and embryonic stem cells as TE insertions in these cells can potentially be transmitted to the next generation.

Regarding the transcriptional repression, the Krüpell-associated box domain-containing zinc-finger proteins (KRAB-ZFPs) present an important role in repressing TE's expression [50]. These proteins comprehend the largest family of transcription factors in humans, containing around 300 different members. KRAB-ZFPs are distinguished by the presence of a zinc-finger domain, which is characterized by an array of C2H2 zinc-fingers that are capable of recognizing specific DNA motifs and conferring binding specificity to these proteins [51]; and a KRAB domain, which mediates the recruitment of repressors that functions as a scaffold for chromatin remodeling complexes, such as TRIM28 (tripartite motif-containing protein 28, also known as KAP1). Both domains point these proteins as mediators of transcriptional repression through epigenetic remodeling. The epigenetic regulators that interact with the KRAB-ZFPs include histone methyltransferases and deacetylases, proteins capable of catalyzing heterochromatin formation, and DNA methyltransferases [50–53]. Together, these proteins are able to impart a repressed state to regions of the chromatin containing TEs. This chromatin remodeling can be more plastic (through histone modifications) or more permanent (through DNA methylation), depending on the TE age: young LTRs are more prone to be silenced by DNA methylation, whereas higher levels of H3K9me3 and H3K9me2 marks were found in intermediate-age LTRs. The evolutionarily old LTRs are more likely inactivated by the accumulation of loss-of-function genetic mutations [54].

The TE regulation mediated by the KRAB-ZFPs plays a key role during the epigenetic remodeling in early embryos. During early development, embryo cells have to clear off methylation markers for the establishment of sex-specific epigenetic markers, and, additionally, most of the previous markers from the gametes, reaching a totipotent state, maintaining only markers on imprinted sites [55–57]. During this wave of demethylation, TEs that were silenced may become active, which can cause germline alterations that will be passed onto future cells' generation. KRAB-ZFPs work by recognizing and repressing TE's expression in the DNA, as well as in the maintenance of imprinted regions during this stage [58, 59].

It is theorized that KRAB-ZFPs have co-evolved with the TEs, in an "arms race model". As an inheritance of this process, KRAB-ZFPs are capable not only of repressing TEs but also TE-derived regulatory sequences [60]. In accordance, a recent study has shown that the overexpression of *ZNF611*, a KRAB-ZFP identified to target *SVA* sequences, was correlated with the enhancing of repression markers in numerous *SVA*s and the reduction of *SVA*-close genes and transpochimeric transcripts levels (transcripts containing fusions of *SVA*s and gene-derived RNAs). Importantly, most of the genes repressed by ZNF611 did not present significant differences in the chromatin marks at their transcription start sites, suggesting that their expression is regulated by *SVA*-derived enhancers controlled by ZNF611 [61].

Whenever a young and active TE escapes from the transcriptional control mediated by KRAB-ZFPs, alternative mechanisms take place to regulate these elements and prevent deleterious consequences. In this context, the innate immune system takes an important part in the post-transcriptional control. The APOBEC (Apolipoprotein B mRNA Editing Catalytic polypeptide like) protein family acts catalyzing the deamination of cytosine to uracil during the reverse transcription of retrovirus, as well as LTR and non-LTR (*L1* and *Alu*) TEs, causing direct degradation or inactivation due to the hypermutation of the complementary DNA (cDNA) [62, 63]. Similarly to the KRAB-ZFP gene family expansion in primate host genomes, which was possibly shaped by the evolution of TEs, an expansion of *APOBEC3 is observed* in New World Monkeys (NWM) [64, 65], followed by the evidence of higher *L1* activity in NWM than in Old World Monkeys and other primates, including humans, through the observations of a higher number of retrocopies in the first [66]. Another protein belonging to the innate immune response that plays a key role in the post-transcriptional regulation of TEs is the adenosine deaminase acting on RNA (ADAR). This enzyme catalyzes the deamination of adenosine to inosine within a double-stranded RNA target and was shown to restrict *L1* retrotransposition. Interestingly, Orecchini and colleagues (2017) suggest that the mechanism by which ADAR1 inhibits *L1* retrotransposition is editing-independent and may be involved in the impairment of the reverse transcriptase step [67].

Another level of TEs repression is based on interactions between the TEs, PIWI proteins and PIWI-interacting RNAs (piRNAs). TEs can be silenced by PIWIs-piRNAs through two different pathways: (1) at the post-transcriptional level, they can cause TE transcript direct degradation through the formation of double-stranded RNAs that are cleaved into small-interfering RNAs (siRNAs), being recognized and degraded by the RNA-induced silencing complex (RISC) [68]; (2) at the transcriptional repression level, nascent LINE-derived transcripts are recognized by PIWIL4 protein/piRNAs complex that recruits methyltransferases (DNMT3L and DNMT3A), leading to *de novo* DNA methylation [69]. piRNAs also contributes to histone modifications, affecting the H3K9me3 methylation state [70], by forming complexes with H1/H3K9me3 and HP1, catalyzing heterochromatin formation [71] in TE-containing regions. PIWI proteins and the associated piRNAs are predominantly expressed in the gonads to protect germ-line genomes from transposable elements [68, 69, 72]. These and other mechanisms of how the host can detect and respond to transposable element activation have been also detailed by Goodier [73].

8.3 Beneficial Impacts of Transposable Elements

Although deleterious impacts may arise from TE insertions, these can be alleviated depending on genome dynamics. In whole-genome duplication events, TE insertions are associated with genome inflation, with an increase in transposition events. At the same time, a relaxed purifying selection tends to happen in these cases, that is, mutations derived from these transposition events tend to be less deleterious and consequently suffer less selective pressure [74]. Insertions that have little or no effects on genome function are better tolerated and may be fixed during evolution. Over evolutionary time, TEs insertions influence genomic plasticity introducing new transcriptional regulators to different *loci*, which can lead to adaptive advantages. Indeed, TE sequences have been repeatedly co-opted for gene regulatory network functions, as can be seen by the fine regulation of numerous ERV-derived transcripts during early embryo differentiation [51, 75, 76]. This abundance of TE-derived insertions in a variety of genomic regions gave rise to many different effects, such as the addition of promoters, insulators, enhancers, and other regulatory elements, which act regulating genome structural organization, and acting as a source of variability by transposing host genes (Fig. 8.1). Each of these roles will be further discussed below.

In a recent study, researchers have shown that newly evolved *cis*-regulatory elements are enriched in young TEs, including LTRs and *SVA*s, functioning as either transcriptional activators or repressors [77]. As previously discussed, TEs can contain RNA polymerase promoters within their structure (Pol II for DNA transposons, HERVs, *L1*, and *SVA*s or Pol III in the case of *Alu* sequences). Therefore, alternative

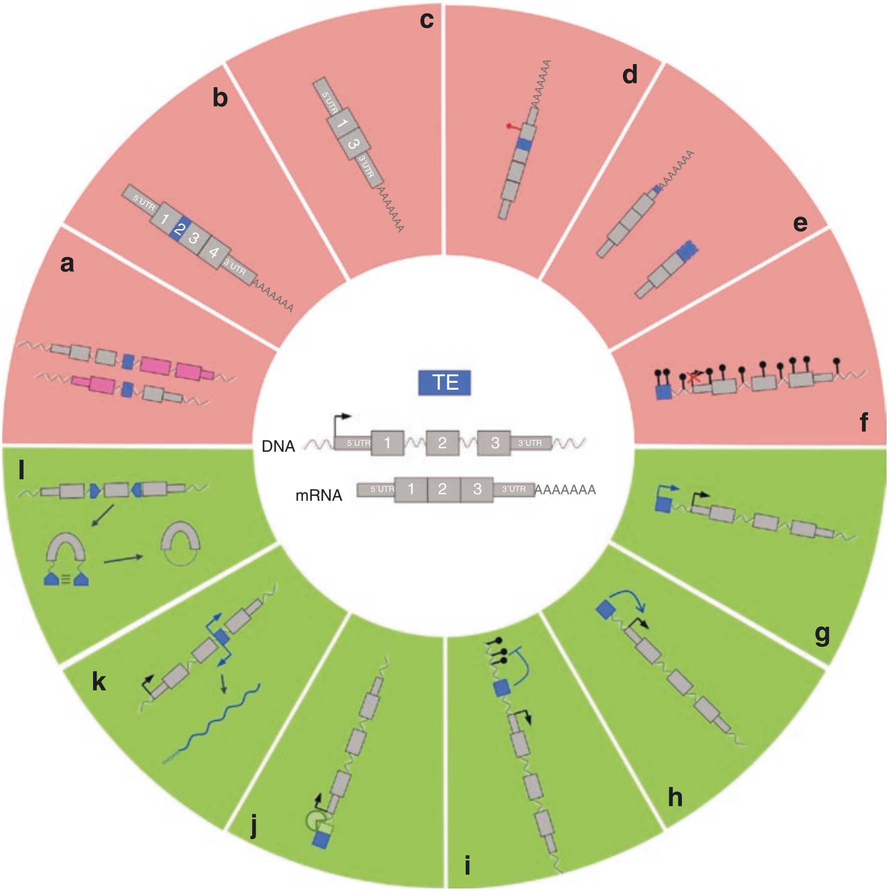


Fig. 8.1 Schematic examples of TEs' impact on the human genome. TEs can negatively (pink background, top panels, a–f) impact a gene/genome in different ways. For example, even long after mobilization, the interaction of TEs at different genomic positions can result in (**a**) genomic rearrangement events—non-allelic homologous recombination. Upon insertion within a gene sequence, a TE can affect the interaction of pre-mRNA and RNA-binding proteins in the splicing process, causing the addition of new exons—exonization (**b**) or subtraction of canonical exons—intronization (**c**). TE insertion inside the coding region can cause frameshift resulting in the presence of a premature stop codon (**d**) and insertions in 3′ UTR can result in shorter mRNA due to alternative poly-adenylation signals or inefficient transcriptional elongation (**e**). Insertions nearby genes can alter its expression through epigenetic silencing of TE surroundings (**f**). On the other hand, TEs can take part in the (*cis* or *trans*) regulatory network of a host genome resulting in positive (green background, bottom panels, g–l) effects. As *cis*-regulatory elements, a new TE insertion can act as an alternative promoter (**g**), as an enhancer (**h**) or an insulator (**i**). Also, TEs can spread transcription factor binding sites (TFBSs) via transduction (**j**). As *trans*-regulatory elements, TEs can affect gene expression through the production of non-coding RNAs. Intronic insertions can generate antisense transcripts that can act as long non-coding RNAs (**k**). Also, the pairing of two inverted TEs can result in the formation of circular RNAs (**l**)

promoters can be derived from the exaptation of a new or remnant insertion of a TE into or near host genes, driving the transcription of this region, thus eliciting gene's expression in new cell types or contexts, establishing novel *cis*-regulatory circuits or fine-tuning a pre-existing network. The use of TE's promoter can also generate a new gene transcript through exonization [78].

Another intriguing evidence of TEs' domestication and exaptation into new cellular functions is associated with the role of primate-specific ERVs and ERV-derived lncRNAs in the regulatory network of early embryo pluripotency during human preimplantation development [79–81]. Human two-cell embryos with knocked down expression of three ERV-derived lncRNAs (*HPAT2*, *HPAT3*, and *HPAT5*) were no longer capable of contributing to the inner cell mass of the blastocyst [82]. In addition to the germline and early embryo TE activity, a TE somatic expression has been observed in the mammalian brain, where *L1* activity is related to altered diversity and complexity of neuronal cell populations [83, 84].

Moreover, TEs can take part in spreading transcription factor binding sites (TFBSs) and act as enhancers in new regulatory networks by their ability to carry the downstream or the upstream flanking sequences during retrotransposition, a process called transduction [85]. As a consequence, some TEs have been shown to contain functional TFBSs [86, 87], which spread through the genome by transposition. However, there is a debate on the actual use of TFBSs that originated from TEs. Although TFBSs are equally distributed in both non-TE and TE-derived regions in the human genome [76], there is a discrepancy in their occupancy: it is estimated that 7–16% of active TFBSs are located within TE-derived sequences [33, 76]. Although less common, the TE-derived TFBSs can be co-opted when the epigenetic suppression is halted during embryonic stem cell differentiation [76]. Another suggested mechanism by which TEs can take part in the host regulatory network is by promoting the emergence of functional elements, specially enhancers, through the insertion of new DNA CpG methylation sequences. However, due to the deamination process that occurs over time, in which 5-methylcytosine is converted to thymine, older TEs tend to show a depletion of CpG islands and an enrichment of mutations (see Chap. 4) [76].

In addition, TEs also make substantial contributions to non-coding regulatory functions that regulate gene expression post-transcriptionally (Fig. 8.1). Notably, 75–83% of human lncRNA transcripts were found to contain TE sequences [88, 89]. Moreover, in some lncRNAs, the pairing of inverted oriented *Alu* elements within intronic regions can mediate RNA circularization (see Chap. 5). The resulting circular RNAs are intrinsically related to a gene expression network involving miRNAs and mRNAs, acting as miRNA sponges, regulators of translation, inductors of alternative splicing, among others [90, 91]. There is also evidence of TE-derived small RNAs acting in gene expression regulation. A recent work by Petri and colleagues (2019) has shown the contribution of *L2* elements as a source of both functional miRNAs and target genes carrying *L2*-derived sequences in their 3′ UTR in the human genome [92].

Beyond their local activity regulating host gene expression, TEs may be important contributors to a broader genomic organization, functioning as insulators and controlling regions of active transcription of large chromosomal regions containing many genes. Insulators can block the interaction between an enhancer and a promoter, functioning as a barrier to prevent the spreading of heterochromatin, and acting as an anchor that assembles chromatin into loops (see Chap. 4). They also harbor domains within which regulatory elements can interact. A recent work by Diehl and colleagues (2020) describes the distribution of TE-derived binding sites for CTCF, a protein involved in the boundaries' definition for chromatin loops in murine and human genomes. They have described that ~35% of CTCF binding sites are derived from TEs, and that TE amplification has had an impact on the 3D genomic landscape, with evidence that TE activity may affect chromatin loop configurations. It was also observed that the looping variability induced by TEs is a major contributor to differential gene expression patterns observed in different cell types or in response to environmental factors [93].

In addition to the already mentioned mechanism of transduction, another source of genetic novelty can be found in *L1*-derived reverse transcriptases, which can perform an RNA-mediated retrotransposition of host genes [94]. It is estimated that there are up to 18,700 gene retrocopies in the human genome, from which ~1,300 are transcribed [95, 96]. These gene duplication events are a rich source for the evolution of adaptive traits, as, after duplication, novel retrocopies may follow one of three evolutionary scenarios: neofunctionalization (evolution of a new biological function or target a novel cellular localization), subfunctionalization (partitioning of biological functions between a gene copy and its parental gene), or conservation of the parental gene function [64].

8.4 Deleterious Impacts of Transposable Elements

Nonetheless, TEs can escape the host defense, be expressed, and cause new insertions, generating deleterious mutations and transcriptional interference on host genes. Not surprisingly, the dysregulated activity of TEs has been linked with several diseases, including cancer, which will be further detailed [37, 97–100]. TEs insertions can disrupt the gene function in several ways, such as: shortening the transcript by incorporating a TE-derived polyadenylation site in the 3′ end of genes [101]; by an inefficient transcriptional elongation through the AT-rich *L1* sequence [102]; altering splicing by the insertion of TE-derived splice acceptor or donor sites, causing intronization (exon skipping) or exonization (creation of new exons), mostly causing a frameshift or creating premature stop codons [101]; or simply altering the epigenetic landscape in the vicinities of TE insertions by increasing local levels of DNA methylation, thus affecting the expression of genes surrounding the insertion (Fig. 8.1) [103]. Moreover, TEs can increase the potential for genetic instability through a process called non-allelic homologous recombination (NAHR) that happens between repeated copies of a TE located at distant genomic positions, contributing to an array of genetic rearrangements, from small-scale deletions and duplications, to chromosomal inversions and translocations (Fig. 8.1) [104, 105]. In addition, it was recently shown that human retrotransposons, in the germline, can drive chromothripsis, a mutational phenomenon that results in genomic rearrangements and is a major process that drives genome evolution in human cancer [106, 107].

TEs have shaped genome evolution in multiple ways. They have also evolved many complex mechanisms and biochemical functions, sometimes acting as a mutagenic element or affecting gene regulatory networks and gene function at either the RNA or DNA level. Misregulated or active elements that evade the above mentioned controlling cellular mechanisms can have a detrimental impact on cellular homeostasis. We will see below evidence showing the implication of TEs on cancer development and response to treatment.

8.5 Transposable Elements and Cancer

Various diseases have been associated with TE activity, such as hemophilia, neuropsychiatric disorders (e.g. schizophrenia, bipolar disorder), neurodevelopmental disorders (e.g. autism), and neurodegenerative disorders (e.g. Parkinson's disease), with a number of reviews and book chapters dedicated to them [108–110]. In this chapter, we will focus on the deleterious impact of these elements associated with cancer.

Although we generally talk about cancer as a unified disease, it is actually an umbrella term and comprises different diseases affecting distinct tissues and organs in our body, each with its particular characteristics. Despite that, all the different cancer types are characterized by uncontrolled cell proliferation, replicative immortalization, ability to evade growth suppressors and the immune system, resistance to cell death, induction of a metabolic shift, and activation of invasion and metastasis [111]. The malignant characteristics are acquired through the acquisition of genomic and epigenomic alterations leading to the activation of oncogenes and inactivation of tumor suppressor genes. The driver alterations in key genes lead cells down the path of malignancy by impacting some of the hallmarks of cancer [111]. We will present examples of the association of TEs with this phenomenon and their impact on cancer development and response to treatment (summarized in Table 8.2).

The number of cancer types and their complexities are enormous, and many are still poorly understood with few pieces of information on how the malignant cells arise from healthy tissue. TE activity plays a role in this transformation, as will be further discussed. We will discuss how TE activity may be established in the malignant cell, and then give examples of how their activity influences the tumor hallmarks.

Aberrant methylation is a phenomenon shared by many tumors. It promotes genomic instability, which may lead to genome rearrangements, copy number variations, and mutations. Different tumor types show distinct gene promoter methylation patterns, which simultaneously change their gene expression profile, impacting important pathways in cancer. This is accompanied by a global genomic hypomethylation in intergenic and repetitive sequences, which frequently include TEs, during cancer initiation and progression, what may elicit the dynamic activation of retrotransposons, especially *L1* elements. Once active, TEs can insert in different regions and, consequently, cause genomic instability, corroborating the concept of cancer as a genomic disorder (Fig. 8.2). *L1* hypomethylation has been linked to poor prognosis in many cancer types,

Table 8.2 Transposable elements and their impact on cancer development and treatment response. Studies and their key findings concerning the impact of TEs on cancer development and therapy response are summarized here

TE	Key findings	Cancer type(s)[a]	References
Alu	Intronic *Alu* on *BRCA1* gene leads to mutant protein without BRCT domain	BRCA	[112]
Alu	*Alu*-derived *BRCA1* rearrangements	BRCA, OV	[113–118]
Alu	Onco-exaptation of TEs	BLCA, BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA and UCEC	[119]
ERV	Onco-exaptation of TEs	BLCA, BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA and UCEC	[119]
		AML	[120]
ERV	5-AZA-CdR induced demethylation leading to expression of ERVs, viral infection mimicry and immune response induction	CRC	[121]
ERV	Ablation of LSD1 leads to ERV expression, viral infection mimicry leading to T-cell anti-tumor response	BRCA, SKCM, LUAD	[122]
ERV	ERV expression leading to a pro-tumorigenic environment (tumors with mesenchymal characteristics)	SCLG	[123]
LINE	*L1* somatic insertion in *the APC* tumor suppressor gene	CRC	[124]
LINE	Onco-exaptation of TEs	BLCA, BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA and UCEC	[119]
LINE	Somatic *L1HS* element insertion in exon 6 of the tumor suppressor gene *PTEN*	UCEC	[125]
LINE	*De novo* insertion of full length LINE-1 element on intron 14 of *RB1*	Familial retinoblastoma	[126]
LINE	*L1* activation after hypomethylation inducing metastasis	SKCM	[127]
LINE	*L1*-induced activation of telomere maintenance genes leading to cell survival, epithelial-mesenchymal transition, invasion and migration	Telomerase-positive tumor cell lines (COAD, SKCM)	[128]

(continued)

Table 8.2 (continued)

TE	Key findings	Cancer type(s)[a]	References
LTR	Onco-exaptation of TEs	AML	[120]
		BLCA, BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA and UCEC	[119]
Mariner/ Tc1 (Tigger3)	Onco-exaptation of TEs	BLCA, BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA and UCEC	[119]

[a] *AML* acute myeloid leukemia, *BLCA* bladder urothelial carcinoma, *BRCA* breast invasive carcinoma, *COAD* colon adenocarcinoma, *CRC* colorectal cancer, *HNSC* head and neck squamous cell carcinoma, *KIRC* kidney renal clear cell carcinoma, *LGG* low-grade glioma, *LIHC* liver hepatocellular carcinoma, *LUAD* lung adenocarcinoma, *LUSC* lung squamous cell carcinoma, *OV* ovarian serous cystadenocarcinoma, *PRAD* prostate adenocarcinoma, *SCLG* small-cell lung cancer, *SKCM* skin cutaneous melanoma, *STAD* stomach adenocarcinoma, *THCA* thyroid carcinoma, *UCEC* uterine corpus endometrial carcinoma

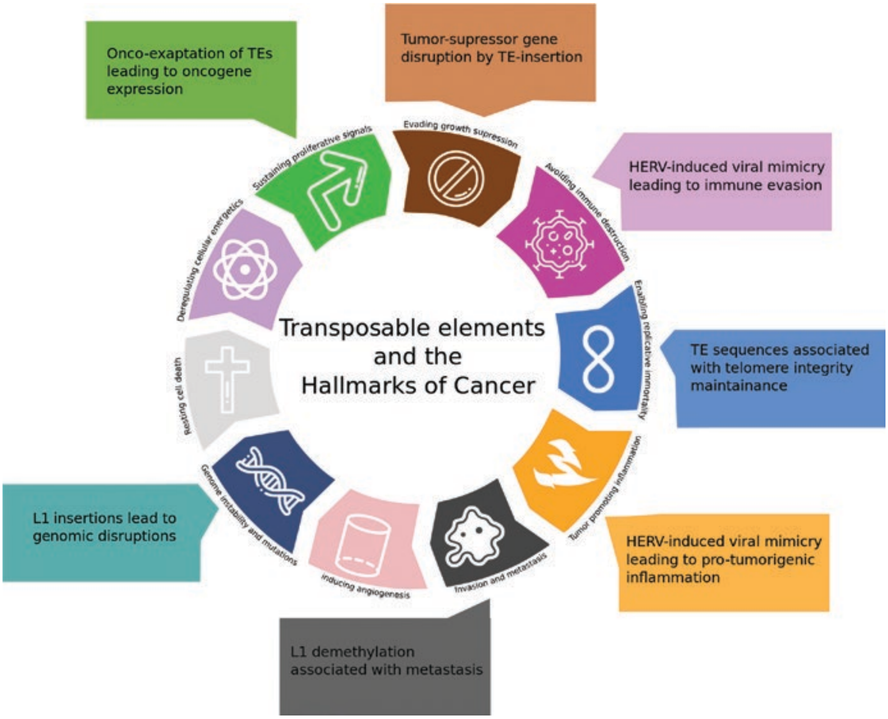

Onco-exaptation of TEs leading to oncogene expression
Tumor-supressor gene disruption by TE-insertion
HERV-induced viral mimicry leading to immune evasion
TE sequences associated with telomere integrity maintainance
HERV-induced viral mimicry leading to pro-tumorigenic inflammation
L1 demethylation associated with metastasis
L1 insertions lead to genomic disruptions
Sustaining proliferative signals
Evading growth supression
Avoiding immune destruction
Enabling replicative immortality
Tumor promoting inflammation
Invasion and metastasis
Inducing angiogenesis
Genome instability and mutations
Resting cell death
Deregulating cellular energetics
Transposable elements and the Hallmarks of Cancer

Fig. 8.2 Impact of TEs on the Hallmarks of Cancer. The image shows the ten hallmarks of cancer according to [111] and highlights how they can be impacted by the TE activity

including breast, colorectal, esophageal, lung, hepatocellular, and gastric [129]. Moreover, *L1* activation due to hypomethylation has also been associated with higher metastatic capacity, with a direct relationship between *L1* demethylation and metastasis events in the five-year follow-up of patients with primary melanoma [127] (Fig. 8.2).

The *L1* retrotransposition during embryonic development, resulting in *L1*-mediated somatic mosaicism, can also favor tumor initiation, through inherited *L1* insertions leading to germline or somatic mutations later on (Fig. 8.2) [130]. Genomic mosaicism generates a variety of cell populations, some of which can function normally or can be prone to become malignant cells, accumulating alterations, and developing genomic instability over years, promoting the tumor onset. Some examples of cancers associated with mosaicism include child leukemias and solid tumors localized in tissues that are under the "field cancerization" effect, such as the esophagus and colorectal cancers [131].

TEs are also involved in telomere maintenance and regulation. This function also impacts cancer progression, as it may enable a replicative immortalized state on malignant cells. *LTR10, MER61, and Alu*-associated sequences have been identified in subtelomeres. They are p53 binding sites involved in telomere lengthening regulation [132]. Another TE also involved in telomere regulation is *L1.* Its activation seems to facilitate the expression of components of the shelterin protein complex, such as KLF-4, c-myc and hTERT (see Chap. 7), which maintain the telomeres' integrity and leads to increased cell survival, contributing to an immortalized state (Fig. 8.2). The proteins coded by the L1-activated genes also seem to induce the expression of SNAIL1, TGF-β, among others, resulting in an epithelial-mesenchymal phenotype transition, which increases invasion and migration of cancer cells (Fig. 8.2; reviewed in [128]).

Regarding driver alterations caused by TE's insertion in key genes associated with cancer (Fig. 8.2), our knowledge is still very incipient. According to Knudson's hypothesis, also known as the two-hit hypothesis, a tumor suppressor gene needs to have both copies inactivated for its function to be completely lost. This implies that the preexistence of a faulty copy of a tumor suppressor gene increases the probability of tumor development [133]. A good example of how this works is the difference regarding the prevalence and age-of-onset distribution between sporadic and familial retinoblastoma cases, wherein the latter, the patient is already born with a damaged copy of the *RB1* (Retinoblastoma transcriptional repressor 1) gene [134]. A recent study has described for the first time the occurrence of TE insertion on the *RB1* gene leading to a case of familial retinoblastoma, showing that the insertion of a *L1* element in the gene leads to erroneous splicing due to the generation of new acceptor splice site [126].

Miki and colleagues, in 1992, were the first to report the disruption of a suppressor gene by the insertion of a mobile element. They showed that the insertion of *L1* in *APC* (Adenomatous Polyposis Coli regulator of WNT signaling pathway), a tumor suppressor gene implicated on the development of colorectal cancer, could be found on colorectal cancer cases [135]. But, due to the lack of technology at the time, it was not possible to identify the origin of the *L1* retrotransposon or prove its role in cancer initiation. This effect, however, has been identified in a later study, which showed the *L1* insertion in the *APC* gene derived from a hot *L1* retrotransposon on chromosome 17 that evaded silencing in the normal tissue. The *L1* insertion was identified in a patient's microsatellite stable tumor, and the *L1* retrotransposon expression was also detected in his healthy tissue. As it was detected in normal and tumor samples, it can be inferred that this insertion could have happened in preneoplastic colonic epithelium, generating an early driver mutation event that may have contributed to the development of colorectal cancer (Fig. 8.2) [124].

As previously stated, *Alu* repeats can create regions susceptible to incorrect chromosome recombinations, which in turn can originate mutations, deletions, and duplications of genes (Fig. 8.2). Examples are deletions in the *BRCA1/2* (Breast Cancer DNA repair associated 1 and 2) genes implicated in cases of familial ovarian and/or breast cancer [136]. Curious about the fact that *BRCA1* germline pathogenic variants were not found as frequently as expected from linkage studies' data, Montagna and collaborators (1999) decided to look at other alterations not captured by common tests (used at the time in clinic) that could be the culprits in the cases with no mutation associated with the disease. The group identified a 3-kb deletion encompassing *BRCA1* exon 17 caused by *Alu*-enabled rearrangement. This alteration caused a frameshift in the protein-coding sequence and created a premature stop codon in two different Italian families with inherited cases of ovarian and breast cancer [137]. Since then, with the advance of technology, various rearrangements facilitated by the abundance of *Alu* sequences in *BRCA1/2* were reported in different populations [113–118].

Genomic instability favors tumorigenesis and cancer progression, but if left uncontrolled, may result in an unsustainable state that will lead to cell death. A mechanism of TE repression that helps to hinder their impact on genomic instability in cancer cells involves the spermatogenic transposon silencer maelstrom (Mael) piRNA-processing factor. Mael is a germline-specific protein that is activated in somatic cells during tumorigenesis that protects cancer cells from spontaneous DNA damage. Kim and collaborators (2016) have demonstrated that Mael depletion in cancer cell lines resulted in the ATM serine/threonine kinase-dependent DNA damage, with an increase in reactive oxygen-species, senescence, and apoptosis in cancer cells [138].

TE expression can also have an impact on the tumor's immunogenicity, influencing the recruitment and activation of immune cells in the tumor microenvironment (Fig. 8.2). Recent studies have shown TE activity inducing the immune response against tumors, but also promoting tumor progression by allowing them to evade from the immune system. Recently, it has been shown that tumors express over 400 TE subfamilies, including HERVs, LINEs, SINEs, and *SVA*s. It was also observed that tumor cells present potentially TE-derived immunogenic peptides, leading to innate immune activation in the tumor, besides contributing to the adaptive immune infiltration, by providing tumor cell surface antigens [139]. However, the immune activation does not always lead to better prognosis. A recent work by Zhu and colleagues (2020) has shown that high TE expression in colorectal cancer (CRC) leads to an "immune overdrive", with the presence of a highly pro-inflammatory infiltrate, resulting in the inflammation state which is another cancer hallmark (Fig. 8.2). They evaluated multiple TE families in CRC samples, correlating the TEs expression with patient survival and immune activation gene pathways. Then, those TEs' expressions were used to generate a score, and the CRC samples were clustered in four groups of increasing risk. High-risk scores were independent of microsatellite instability status (commonly associated with global hypomethylation in CRC) or mutation burden, and predicted worse prognosis, showing that TEs expression could be used as a biomarker [140]. As another example of immune activation causing a deleterious effect, it has been shown that the expression of HERVs may trigger signaling cascades that ultimately lead to a pro-tumorigenic environment. For example, in small-cell lung cancer, the expression of a subset of HERVs called SPARCS (Stimulated 3 prime antisense retroviral coding sequences) leads to the production of double-stranded RNA (dsRNA), mimicking a viral infection. In turn, it initiated a response cascade that signals for the release of cytokines and the expression of *PD-L1* by the subgroup of mesenchymal tumor cells when exposed to IFN-γ. In this case, despite the immune cell infiltration in the tumor microenvironment, it presents an immunosuppressive profile, impairing the antiviral immune response expected to be activated, the MDA5/MAVS/IRF7 pathway, that enables the clearing of tumor cells [123].

As explained before, TEs can also add new expression regulator elements and their use can lead to the aberrant expression of oncogenes, a process that was termed onco-exaptation by Babain and Mager (Fig. 8.2) [141]. An example of this use of TEs in cancer was reported for acute myeloid leukemia (AML). AML is a cancer of the myeloid line of blood cells, with a great diversity of cytogenetic abnormalities. It is characterized by the clonal expansion of cells from the hematopoietic system that are abnormally or poorly differentiated [142]. Malignant cells in AML present diversification in its epigenetic and genetic landscapes, leading to high variability in the disease intra- and inter-patient [143]. Deniz and colleagues (2020) observed that some of the genetic variability can be attributed to the onco-exaptation of TE sequences as enhancers for genes implicated in hematopoiesis and AML pathogenesis (Fig. 8.2). They identified ERVs associated with AML and CD34+ hematopoietic stem cell-specific DNAse-hypersensitive sites (DHS) that had histone markers for active enhancers in commonly used laboratory cell lines with different genetic and cytogenetic backgrounds, as well as in AML samples. Some of these potential active enhancer regions presented LTR sequences from ERVs and were enriched for AML-related TF binding motifs. The deletion by CRISPR-Cas9 of candidate ERV regions led to differences in gene expression and cell growth suppression. Their findings point to these regions' possible exploitation by cancer cells to promote cell proliferation, survival, and maintenance of a dedifferentiated cell state (Fig. 8.2) [120].

A recent study analyzed tumor and tumor-matched-normal samples (7,769 and 625 samples, respectively) from the "The Cancer Genome Atlas" (TCGA) project, corresponding to 15 different cancer types, to identify onco-exaptation events in multiple tumors all at once. They were able to identify events enriched in specific tumors and also events that are common to multiple tumors [119]. A total of 625 TE-oncogene chimeric transcripts were identified and approximately half of all the tumors analyzed had at least one instance of onco-exaptation event identified (Fig. 8.2). The prevalence of events across cancer types ranged considerably suggesting that, although TE expression in cancer is often tumor-specific, onco-exaptation can be a promiscuous mechanism for the activation of oncogenes. However, how this phenomenon occurs and the extent to which it contributes to oncogenesis remain unknown.

8.6 Transposable Elements Activation and Cancer Treatment

Cancer treatment can vary from a more aggressive and systemic approach to the more localized targeted therapies, depending on the type of cancer and its characteristics. Treatment responses vary according to each patient and may be influenced by alterations that arise from TEs activation. Besides influencing the treatment response, TEs activation can also be used as a target for new therapeutic strategies, as we will discuss further.

As discussed previously, the presence of *Alu* repeats on the *BRCA* genes can lead to mutations and/or deletions associated with cancer. Although patients with *BRCA1* germline mutations have a higher cumulative risk of developing breast and ovarian cancer than the general population [144], germline or somatic mutations in this gene, independently of origin, are also associated with better survival outcomes and therapy response [145]. It is important to mention that carcinomas harboring *BRCA1* mutations are generally sensitive to PARP inhibitor (PARPi) therapy [146]. However, in a recent study, intronic *Alus* on the *BRCA1* gene were shown to produce a mutant protein that did not contain the BRCT domain. A BRCA1 mutant protein containing the BRCT domain does not fold correctly and is, therefore, targeted for degradation. The protein that arises from this BRCT domain-deficient isoform avoids its degradation and is still functional in the DNA repair pathway, which in turn promotes PARPi resistance [112] (Fig. 8.3). The repression of TEs was also associated with taxane resistance in triple-negative breast cancer cells. Deblois and collaborators (2020) have shown that these cells undergo metabolic adaptations, which result in regions with a high rate of histone H3 lysine methylation, leading to TE repression. They have found that inhibiting the *EZH2* H3K27me3 methyltransferase leads to an overall TE-derived sequence expression, with an accumulation of dsRNA fragments and the activation of the viral mimicry response, showing the potential use of this mechanism for treating chemoresistant breast cancer [147] (Fig. 8.3).

Besides promoting treatment resistance, the expression of TEs can also favor an anti-tumor response. As TEs are silenced by methylation, the use of demethylating agents can lead to their expression. A study by Roulois and collaborators showed that when treated with 5-aza-2-deoxycytidine (5-AZA-CdR), colorectal cancer cell lines started expressing HERVs. 5-AZA-CdR is a cytidine analog, a demethylating agent used to induce global demethylation by trapping the proteins from the DNA-methylation machinery and recovering tumor suppressor genes expression. Roulois and collaborators (2015) identified the formation of dsRNA derived from HERVs on the cell lines upon treatment. By mimicking a viral infection, the aforementioned antiviral immune response is activated, enabling the elimination of tumor cells [121].

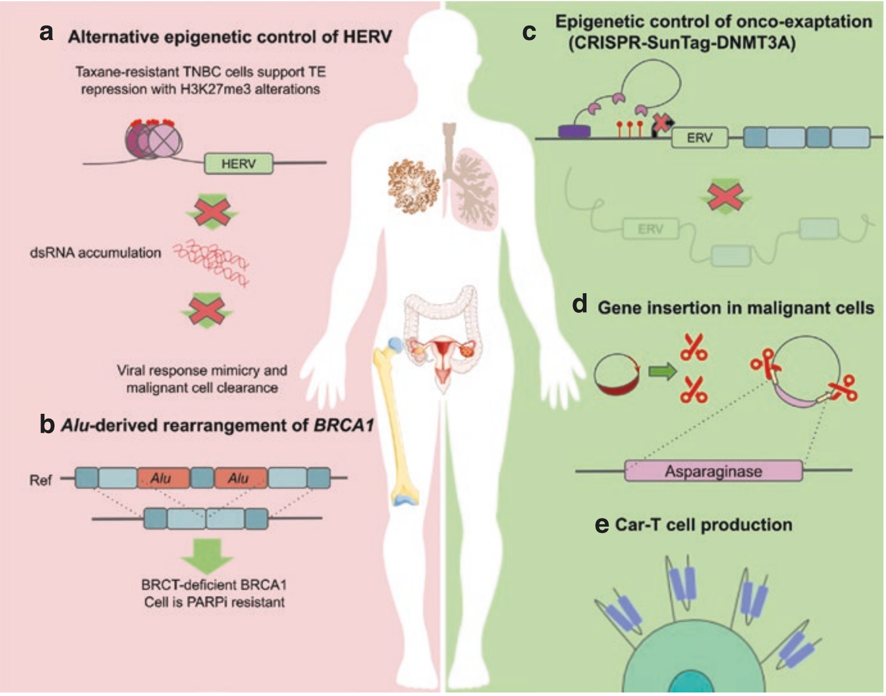

a Alternative epigenetic control of HERV
Taxane-resistant TNBC cells support TE repression with H3K27me3 alterations
HERV
dsRNA accumulation
Viral response mimicry and malignant cell clearance
b *Alu*-derived rearrangement of *BRCA1*
Ref
Alu
Alu
BRCT-deficient BRCA1
Cell is PARPi resistant
c Epigenetic control of onco-exaptation (CRISPR-SunTag-DNMT3A)
ERV
ERV
d Gene insertion in malignant cells
Asparaginase
e Car-T cell production

Fig. 8.3 Harmful or helpful for cancer? On the left side (pink panel), there are examples of ways by which TEs can interfere on cancer treatment response. (**a**) On taxane-resistant triple-negative breast cancer (TNBC) cells, metabolic stress leads to a repression of HERV expression, compensating the hypomethylation of those elements by augmenting the presence of H3K27me3 markers. This leads to the decrease in accumulation of dsRNA and the mimicry of viral response is not triggered, impacting malignant cell clearance by the immune system; (**b**) *Alu*-derived rearrangement of *BRCA1* produces a protein devoid of the BRCT domain, which impacts its targeting for degradation. As a consequence of the persistence of BRCA1, the cell becomes PARPi-resistant. On the right side (green panel), examples of ways to control target TEs or to use TEs as a tool for cancer treatment. (**c**) Guided methylation of TEs by CRISPR-SunTag-DNMT3A for onco-exaptation control; (**d**) insertion of genes coding for enzymes capable of degrading important molecules for malignant cell survival using the Sleeping Beauty transposon delivery system; (**e**) production of T cells capable of recognizing and responding to the tumor by inserting gene sequences of chimeric receptors using the Sleeping Beauty transposon delivery system

A similar effect was observed in a study that showed the effects of the histone demethylase *LSD1* ablation in cancer cells. The LSD1 protein normally represses the expression of ERV*s* and its downstream effects on the cell. In this study, *LSD1* expression was depleted by shRNA knockdown, CRISPR/Cas9 deletion, or by using the GSK-LSD1 catalytic inhibitor. When *LSD1* was not expressed, tumor cells showed cell growth arrest, enrichment of the antigen presentation pathway, and enhancement of tumor infiltration with T cytotoxic cells, promoting T-cell antitumor response. Tumor resistance to PD-1 blockade was also overcome, showing *LSD1* ablation's potential as a complement to immunotherapy [122].

TEs can also be used as targets for the development of new therapeutics for cancer. For instance, it was recently shown that TEs expression generates several antigenic peptides that are conserved among different tumor types. The authors suggested that these potential neoantigens should be evaluated not only as "off-the-shelf" vaccine targets for therapeutic intervention but also for cancer prophylaxis [139]. In a different perspective, Jang and collaborators (2019) identified TE-derived oncogenic transcripts across 15 cancer types showed that in vitro target epigenetic silencing of the *AluJb*-derived *LIN28B* TE, identified in liver cancer and lung cancer cell lines, was capable of reducing the TE-regulated oncogene expression [119] (Fig. 8.3). These results pave the way for the use of guided methylation of TEs as a targeted therapy, in cases where onco-exaptation was identified.

8.7 The Use of Transposable Elements as Tools in Precision Medicine

We have discussed the impact of endogenous TEs on the evolution and diversification of genomes, the deleterious impact of its activity in cancer and treatment response, as well as their use as targets for new therapeutic strategies. Here we will discuss how TEs are being used as tools in already approved therapy and possible new uses in precision medicine in cancer.

One of the main characteristics of malignant cells that successfully proliferate is their capacity to elude the immune response. The immune evasion can occur by either reducing or annulling antigen presentation, by hampering immune response activation, and by subverting immune cells to provide a favorable environment for its growth [148]. To revert this state, immunotherapies based on checkpoint blockade have been developed and shown to be able to recuperate anti-tumor immune response [149]. Modified cell-based immunotherapies known as adoptive cell therapy (ACT) have also shown excellent results [150].

ACT is based on the patient's cells to generate an anti-tumor response. These can be tumor-infiltrating lymphocytes (TILs) or genetically modified T cells expressing novel T cell receptors (TCR) or chimeric antigen receptors (CAR). In particular, CAR-T cells that express a receptor capable of recognizing CD19 have already been approved by the FDA for treatment of refractory B cell lymphomas and have shown positive results [151, 152]. These cells are generally modified using a lentiviral or viral process, which is not only time-consuming but also costly, deeming the treatment almost impossible for the general population [153].

Nevertheless, recently the Sleeping Beauty (SB) transposon-transposase delivery system for the generation of CAR-T cells was developed. The SB transposon is part of the *Tc1/mariner* superfamily which is widespread in eukaryotic organisms. A functional gene encoding the SB transposase was produced by "reverse-evolution" of DNA sequences from different fish species' transposons [154]. This transposase is capable of inserting engineered DNA-transposon sequences flanked by TIRs into

TA-rich regions in the genome [155]. Recent studies have shown that the protocol using the SB delivery system is capable of generating modified cells in a secure, more cost-efficient way in a fraction of the time needed for production using a lentiviral approach, turning these into a viable option for treatment, even in developing countries and underserved public health systems [156]. A phase I clinical trial using SB-developed CAR-T cells in patients with advanced non-Hodgkin lymphomas and acute lymphoblastic leukemia has shown promising results, with 83% progression-free survival in 30 months for patients subjected to autologous hematopoietic stem cell transplantation (HSCT) (Fig. 8.3) [157].

Another use of SB in cancer studies involves the direct modification of the malignant cell. Numerous studies have shown the potential of the SB system for integrating target genes in both cell cultures and mice [158–160]. It has also been used as a tool for screening genes that contribute to cancer development and metastasis [161–164], and for developing carcinogenesis models [165]. Recent studies also show the potential use of this system for direct gene therapy with malignant cells as targets [160], with cytotoxic effects already demonstrated in lung cancer cells (Fig. 8.3) [166].

8.8 Concluding Remarks

There is still a lot to be learned and there are still methodological problems to be tackled to fine-tune the identification of TEs and TE insertions. For cancer, in particular, knowing which alterations arise normally from *de novo* or germline insertions, as well as understanding the role of TE expression in different points of disease progression and different cell states might be valuable. Not only can it help in understanding the disease, but also in developing treatments tailored to this information. Nevertheless, collectively, the data presented in this chapter demonstrate the still evolving knowledge of the TEs' impact not only on the genome content, structure and regulatory hierarchy, but also its implications in development, be it in health or disease conditions. It also highlights the potential of this once called "junk DNA" to be used not only as a tool for research and treatment, but also as a target for new therapies.

Bibliography

1. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36:344–55.
2. McClintock B. Intranuclear systems controlling gene action and mutation. Brookhaven Symp Biol. 1956:58–74.
3. Guio L, González J. New insights on the evolution of genome content: population dynamics of transposable elements in flies and humans. Methods Mol Biol. 2019;1910:505–30.
4. Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. Chromosom Res. 2018;26:25–43.

5. Platt RN, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 2016;8:403–10.
6. Nishihara H. Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. Genes Genet Syst. 2020;94:269–81.
7. Ricci M, Peona V, Guichard E, Taccioli C, Boattini A. Transposable elements activity is positively related to rate of speciation in mammals. J Mol Evol. 2018;86:303–10.
8. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
9. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7:e1002384.
10. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989;5:103–7.
11. Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 2017;18:292–308.
12. Wicker T, Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
13. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008;9:411–2. author reply 414
14. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA. 2017;8:19.
15. Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst. 2020;94:233–52.
16. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
17. Kojima KK. Human transposable elements in Repbase: genomic footprints from fish to humans. Mob DNA. 2018;9:2.
18. Ray DA, Feschotte C, Pagan HJT, Smith JD, Pritham EJ, Arensburger P, Atkinson PW, Craig NL. Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res. 2008;18:717–28.
19. Mitra R, Li X, Kapusta A, Mayhew D, Mitra RD, Feschotte C, Craig NL. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. Proc Natl Acad Sci U S A. 2013;110:234–9.
20. Huang S, Tao X, Yuan S, et al. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. Cell. 2016;166:102–14.
21. Carmona LM, Schatz DG. New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. FEBS J. 2017;284:1590–605.
22. Kapitonov VV, Koonin EV. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biol Direct. 2015;10:20.
23. Pavelitz T, Gray LT, Padilla SL, Bailey AD, Weiner AM. PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. Mob DNA. 2013;4:23.
24. Henssen AG, Henaff E, Jiang E, et al. Genomic DNA transposition induced by human PGBD5. elife. 2015;4:e10565. https://doi.org/10.7554/eLife.10565.
25. Majumdar S, Singh A, Rio DC. The human THAP9 gene encodes an active P-element DNA transposase. Science. 2013;339:446–8.
26. Cappello J, Handelsman K, Lodish HF. Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell. 1985;43:105–15.
27. Malicki M, Spaller T, Winckler T, Hammann C. DIRS retrotransposons amplify via linear, single-stranded cDNA intermediates. Nucleic Acids Res. 2020;48:4230–43.
28. Ishak CA, De Carvalho DD. Reactivation of endogenous retroelements in cancer development and therapy. Annu Rev Cancer Biol. 2020;4:159–76. https://doi.org/10.1146/annurev-cancerbio-030419-033525.

29. Weiss RA. Human endogenous retroviruses: friend or foe? APMIS. 2016;124:4–10.
30. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A. 2016;113:E2326–34.
31. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta. 2012;33:663–71.
32. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351:1083–7.
33. Kellner M, Makałowski W. Transposable elements significantly contributed to the core promoters in the human genome. Sci China Life Sci. 2019;62:489–97.
34. Guichard E, Peona V, Malagoli Tagliazucchi G, et al. Impact of non-LTR retrotransposons in the differentiation and evolution of anatomically modern humans. Mob DNA. 2018;9:28.
35. Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, Ha H, Xing J, Jorde LB. Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res. 2019;29:1567–77.
36. Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet. 2011;12:187–215.
37. Kazazian HH, Moran JV. Mobile DNA in health and disease. N Engl J Med. 2017;377:361–70.
38. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. LINE-1 retrotransposition activity in human genomes. Cell. 2010;141:1159–70.
39. Philippe C, Vargas-Landin DB, Doucet AJ, van Essen D, Vera-Otarola J, Kuciak M, Corbin A, Nigumann P, Cristofari G. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. elife. 2016;5:e13926. https://doi.org/10.7554/eLife.13926.
40. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. Mob DNA. 2016;7:9.
41. Deininger P. Alu elements: know the SINEs. Genome Biol. 2011;12:236.
42. Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J. Evolutionary history of 7SL RNA-derived SINEs in supraprimates. Trends Genet. 2007;23:158–61.
43. Kryatova MS, Steranka JP, Burns KH, Payer LM. Insertion and deletion polymorphisms of the ancient AluS family in the human genome. Mob DNA. 2017;8:6.
44. Konkel MK, Walker JA, Hotard AB, Ranck MC, Fontenot CC, Storer J, Stewart C, Marth GT, 1000 Genomes Consortium, Batzer MA. Sequence analysis and characterization of active human alu subfamilies based on the 1000 genomes pilot project. Genome Biol Evol. 2015;7:2608–22.
45. Zeng L, Pederson SM, Cao D, Qu Z, Hu Z, Adelson DL, Wei C. Genome-wide analysis of the association of transposable elements with gene regulation suggests that Alu elements have the largest overall regulatory impact. J Comput Biol. 2018;25:551–62.
46. Ono M, Kawakami M, Takezawa T. A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res. 1987;15:8725–37.
47. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res. 2012;40:1666–83.
48. Larsen PA. Transposable elements and the multidimensional genome. Chromosom Res. 2018;26:1–3.
49. Gerdes P, Richardson SR, Mager DL, Faulkner GJ. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome Biol. 2016;17:100.
50. Yang P, Wang Y, Macfarlan TS. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. Trends Genet. 2017;33:871–81.
51. Imbeault M, Helleboid P-Y, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature. 2017;543:550–4.
52. Ecco G, Cassano M, Kauzlaric A, Duc J, Coluccio A, Offner S, Imbeault M, Rowe HM, Turelli P, Trono D. transposable elements and their KRAB-ZFP controllers regulate gene expression in adult tissues. Dev Cell. 2016;36:611–23.

53. Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature. 2014;516:242–5.
54. Ohtani H, Liu M, Zhou W, Liang G, Jones PA. Switching roles for DNA and histone methylation depend on evolutionary ages of human endogenous retroviruses. Genome Res. 2018;28:1147–57.
55. Smallwood SA, Kelsey G. De novo DNA methylation: a germ cell perspective. Trends Genet. 2012;28:33–42.
56. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev. 2014;28:812–28.
57. Zeng Y, Chen T. DNA methylation reprogramming during mammalian development. Genes (Basel). 2019;10(4):257. https://doi.org/10.3390/genes10040257.
58. Coluccio A, Ecco G, Duc J, Offner S, Turelli P, Trono D. Individual retrotransposon integrants are differentially controlled by KZFP/KAP1-dependent histone methylation, DNA methylation and TET-mediated hydroxymethylation in naïve embryonic stem cells. Epigenetics Chromatin. 2018;11:7.
59. Jansz N. DNA methylation dynamics at transposable elements in mammals. Essays Biochem. 2019;63:677–89.
60. Helleboid P-Y, Heusel M, Duc J, et al. The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. EMBO J. 2019;38:e101220.
61. Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. Cell Stem Cell. 2019;24:724–735.e5.
62. Knisbacher BA, Levanon EY. DNA editing of LTR retrotransposons reveals the impact of apobecs on vertebrate genomes. Mol Biol Evol. 2016;33:554–67.
63. Orecchini E, Frassinelli L, Galardi S, Ciafrè SA, Michienzi A. Post-transcriptional regulation of LINE-1 retrotransposition by AID/APOBEC and ADAR deaminases. Chromosom Res. 2018;26:45–59.
64. Yang L, Emerman M, Malik HS, McLaughlin RN. Retrocopying expands the functional repertoire of APOBEC3 antiviral proteins in primates. elife. 2020;9:e58436. https://doi.org/10.7554/eLife.58436.
65. Moreira DA, Lamarca AP, Soares RF, Coelho AMA, Furtado C, Scherer NM, Moreira MAM, Seuánez HN, Boroni M. Transcriptome of the Southern Muriqui Brachyteles arachnoides (Primates:Platyrrhini), a critically endangered new world monkey: evidence of adaptive evolution. Front Genet. 2020;11:831.
66. Navarro FCP, Galante PAF. A genome-wide landscape of retrocopies in primate genomes. Genome Biol Evol. 2015;7:2265–75.
67. Orecchini E, Doria M, Antonioni A, Galardi S, Ciafrè SA, Frassinelli L, Mancone C, Montaldo C, Tripodi M, Michienzi A. ADAR1 restricts LINE-1 retrotransposition. Nucleic Acids Res. 2017;45:155–68.
68. Russell SJ, LaMarre J. Transposons and the PIWI pathway: genome defense in gametes and embryos. Reproduction. 2018;156:R111–24.
69. Molaro A, Falciatori I, Hodges E, Aravin AA, Marran K, Rafii S, McCombie WR, Smith AD, Hannon GJ. Two waves of de novo methylation during mouse germ cell development. Genes Dev. 2014;28:1544–9.
70. Pezic D, Manakov SA, Sachidanandam R, Aravin AA. piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. Genes Dev. 2014;28:1410–28.
71. Iwasaki YW, Murano K, Ishizu H, Shibuya A, Iyoda Y, Siomi MC, Siomi H, Saito K. Piwi modulates chromatin accessibility by regulating multiple factors including histone H1 to repress transposons. Mol Cell. 2016;63:408–19.
72. Tóth KF, Pezic D, Stuwe E, Webster A. The piRNA pathway guards the germline genome against transposable elements. Adv Exp Med Biol. 2016;886:51–77.

73. Goodier JL. Restricting retrotransposons: a review. Mob DNA. 2016;7:16.
74. Baduel P, Quadrana L, Hunter B, Bomblies K, Colot V. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat Commun. 2019;10:5818.
75. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18:71–86.
76. Zhou W, Liang G, Molloy PL, Jones PA. DNA methylation enables transposable element-driven genome expansion. Proc Natl Acad Sci U S A. 2020;117:19359–66.
77. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. Transposable elements are the primary source of novelty in primate gene regulation. Genome Res. 2017;27:1623–33.
78. Jung J, Lee S, Cho H-S, Park K, Ryu J-W, Jung M, Kim J, Kim H, Kim D-S. Bioinformatic analysis of regulation of natural antisense transcripts by transposable elements in human mRNA. Genomics. 2019;111:159–66.
79. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42:631–4.
80. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012;487:57–63.
81. Grow EJ, Flynn RA, Chavez SL, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature. 2015;522:221–5.
82. Durruthy-Durruthy J, Sebastiano V, Wossidlo M, et al. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat Genet. 2016;48:44–52.
83. Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci. 2014;15:497–506.
84. Baillie JK, Barnett MW, Upton KR, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479:534–7.
85. Chu C, Zhao B, Park PJ, Lee EA. Identification and genotyping of transposable element insertions from genome sequencing data. Curr Protoc Hum Genet. 2020;107:e102.
86. Testori A, Caizzi L, Cutrupi S, Friard O, De Bortoli M, Cora' D, Caselle M. The role of transposable elements in shaping the combinatorial interaction of transcription factors. BMC Genomics. 2012;13:400.
87. Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philos Trans R Soc Lond Ser B Biol Sci. 2020;375:20190347.
88. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13:R107.
89. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013;9:e1003470.
90. Liang Z-Z, Guo C, Zou M-M, Meng P, Zhang T-T. circRNA-miRNA-mRNA regulatory network in human lung cancer: an update. Cancer Cell Int. 2020;20:173.
91. Zou F-W, Cao D, Tang Y-F, Shu L, Zuo Z, Zhang L-Y. Identification of CircRNA-miRNA-mRNA regulatory network in gastrointestinal stromal tumor. Front Genet. 2020;11:403.
92. Petri R, Brattås PL, Sharma Y, Jönsson ME, Pircs K, Bengzon J, Jakobsson J. LINE-2 transposable elements are a source of functional human microRNAs and target sites. PLoS Genet. 2019;15:e1008036.
93. Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. Nat Commun. 2020;11:1796.
94. Kubiak MR, Makałowska I. Protein-coding genes' retrocopies and their functions. Viruses. 2017;9(4):80. https://doi.org/10.3390/v9040080.

95. Cerbin S, Jiang N. Duplication of host genes by transposable elements. Curr Opin Genet Dev. 2018;49:63–9.
96. Casola C, Betrán E. The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? Genome Biol Evol. 2017;9:1351–73.
97. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17:415–24.
98. Jönsson ME, Garza R, Johansson PA, Jakobsson J. Transposable elements: a common feature of neurodevelopmental and neurodegenerative disorders. Trends Genet. 2020;36:610–23.
99. Belancio VP, Deininger PL, Roy-Engel AM. LINE dancing in the human genome: transposable elements and disease. Genome Med. 2009;1:97.
100. Chenais B. Transposable elements in cancer and other human diseases. Curr Cancer Drug Targets. 2015;15:227–42.
101. Lavi E, Carmel L. Alu exaptation enriches the human transcriptome by introducing new gene ends. RNA Biol. 2018;15:715–25.
102. Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature. 2004;429:268–74.
103. Saze H. Epigenetic regulation of intragenic transposable elements: a two-edged sword. J Biochem. 2018;164:323–8.
104. Underwood CJ, Choi K. Heterogeneous transposable elements as silencers, enhancers and targets of meiotic recombination. Chromosoma. 2019;128:279–96.
105. Roychowdhury T, Abyzov A. Chromatin organization modulates the origin of heritable structural variations in human genome. Nucleic Acids Res. 2019;47:2766–77.
106. Nazaryan-Petersen L, Bertelsen B, Bak M, Jønson L, Tommerup N, Hancks DC, Tümer Z. Germline chromothripsis driven by L1-mediated retrotransposition and Alu/Alu homologous recombination. Hum Mutat. 2016;37:385–95.
107. Cortés-Ciriano I, Lee JJ-K, Xi R, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet. 2020;52:331–41.
108. Guffanti G, Gaudi S, Fallon JH, Sobell J, Potkin SG, Pato C, Macciardi F. Transposable elements and psychiatric disorders. Am J Med Genet B Neuropsychiatr Genet. 2014;165B:201–16.
109. Misiak B, Ricceri L, Sąsiadek MM. Transposable elements and their epigenetic regulation in mental disorders: current evidence in the field. Front Genet. 2019;10:580.
110. Payer LM, Burns KH. Transposable elements in human genetic disease. Nat Rev Genet. 2019;20:760–72.
111. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
112. Wang Y, Bernhardy AJ, Nacson J, et al. BRCA1 intronic Alu elements drive gene rearrangements and PARP inhibitor resistance. Nat Commun. 2019;10:5661.
113. The BRCA1 Exon 13 Duplication Screening Group. The exon 13 duplication in the BRCA1 gene is a founder mutation present in geographically diverse populations. The BRCA1 Exon 13 Duplication Screening Group. Am J Hum Genet. 2000;67:207–12.
114. Kremeyer B, Soller M, Lagerstedt K, Maguire P, Mazoyer S, Nordling M, Wahlström J, Lindblom A. The BRCA1 exon 13 duplication in the Swedish population. Familial Cancer. 2005;4:191–4.
115. Judkins T, Rosenthal E, Arnell C, Burbidge LA, Geary W, Barrus T, Schoenberger J, Trost J, Wenstrup RJ, Roa BB. Clinical significance of large rearrangements in BRCA1 and BRCA2. Cancer. 2012;118:5210–6.
116. Concolino P, Rizza R, Hackmann K, Paris I, Minucci A, De Paolis E, Scambia G, Zuppi C, Schrock E, Capoluongo E. Characterization of a new BRCA1 rearrangement in an Italian woman with hereditary breast and ovarian cancer syndrome. Breast Cancer Res Treat. 2017;164:497–503.
117. Su L, Zhang J, Meng H, Ouyang T, Li J, Wang T, Fan Z, Fan T, Lin B, Xie Y. Prevalence of BRCA1/2 large genomic rearrangements in Chinese women with sporadic triple-negative or familial breast cancer. Clin Genet. 2018;94:165–9.
118. van der Merwe NC, Oosthuizen J, Theron M, Chong G, Foulkes WD. The contribution of large genomic rearrangements in BRCA1 and BRCA2 to South African familial breast cancer. BMC Cancer. 2020;20:391.

119. Jang HS, Shah NM, Du AY, et al. Transposable elements drive widespread expression of oncogenes in human cancers. Nat Genet. 2019;51:611–7.
120. Deniz Ö, Ahmed M, Todd CD, Rio-Machin A, Dawson MA, Branco MR. Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. Nat Commun. 2020;11:3506.
121. Roulois D, Loo Yau H, Singhania R, et al. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. Cell. 2015;162:961–73.
122. Sheng W, LaFleur MW, Nguyen TH, et al. LSD1 ablation stimulates anti-tumor immunity and enables checkpoint blockade. Cell. 2018;174:549–563.e19.
123. Cañadas I, Thummalapalli R, Kim JW, et al. Tumor innate immunity primed by specific interferon-stimulated endogenous retroviruses. Nat Med. 2018;24:1143–50.
124. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. 2016;26:745–55.
125. Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Res. 2014;24:1053–63.
126. Rodríguez-Martín C, Cidre F, Fernández-Teijeiro A, Gómez-Mariano G, de la Vega L, Ramos P, Zaballos Á, Monzón S, Alonso J. Familial retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene. J Hum Genet. 2016;61:463–6.
127. Ecsedi SI, Hernandez-Vargas H, Lima SC, Herceg Z, Adany R, Balazs M. Transposable hypomethylation is associated with metastatic capacity of primary melanomas. Int J Clin Exp Pathol. 2013;6:2943–8.
128. Mueller C, Aschacher T, Wolf B, Bergmann M. A role of LINE-1 in telomere regulation. Front Biosci (Landmark Ed). 2018;23:1310–9.
129. Ponomaryova AA, Rykova EY, Gervas PA, Cherdyntseva NV, Mamedov IZ, Azhikina TL. Aberrant methylation of LINE-1 transposable elements: a search for cancer biomarkers. Cell. 2020;9(9):2017. https://doi.org/10.3390/cells9092017.
130. Faulkner GJ, Garcia-Perez JL. L1 mosaicism in mammals: extent, effects, and evolution. Trends Genet. 2017;33:802–16.
131. Fernández LC, Torres M, Real FX. Somatic mosaicism: on the road to cancer. Nat Rev Cancer. 2016;16:43–55.
132. Lieberman PM. Retrotransposon-derived p53 binding sites enhance telomere maintenance and genome protection. BioEssays. 2016;38:943–9.
133. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971;68:820–3.
134. Kaewkhaw R, Rojanaporn D. Retinoblastoma: etiology, modeling, and treatment. Cancers (Basel). 2020;12(8):2304. https://doi.org/10.3390/cancers12082304.
135. Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. 1992;52:643–5.
136. Ford D, Easton DF, Stratton M, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet. 1998;62:676–89.
137. Montagna M, Santacatterina M, Torri A, Menin C, Zullato D, Chieco-Bianchi L, D'Andrea E. Identification of a 3 kb Alu-mediated BRCA1 gene rearrangement in two breast/ovarian cancer families. Oncogene. 1999;18:4160–5.
138. Kim S-H, Park E-R, Cho E, Jung W-H, Jeon J-Y, Joo H-Y, Lee K-H, Shin H-J. Mael is essential for cancer cell survival and tumorigenesis through protection of genetic integrity. Oncotarget. 2017;8:5026–37.
139. Kong Y, Rose CM, Cass AA, et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. Nat Commun. 2019;10:5228.
140. Zhu X, Fang H, Gladysz K, Barbour JA, Wong JWH. Overexpression of transposable elements underlies immune overdrive and poor clinical outcome in cancer patients. medRxiv. 2020; https://doi.org/10.1101/2020.07.14.20129031.

141. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. Mob DNA. 2016;7:24.
142. Döhner H, Weisdorf DJ, Bloomfield CD. Acute myeloid leukemia. N Engl J Med. 2015;373:1136–52.
143. Li S, Mason CE, Melnick A. Genetic and epigenetic heterogeneity in acute myeloid leukemia. Curr Opin Genet Dev. 2016;36:100–6.
144. Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. JAMA. 2017;317:2402–16.
145. Xu K, Yang S, Zhao Y. Prognostic significance of BRCA mutations in ovarian cancer: an updated systematic review with meta-analysis. Oncotarget. 2017;8:285–302.
146. Faraoni I, Graziani G. Role of BRCA mutations in cancer treatment with poly(ADP-ribose) polymerase (PARP) inhibitors. Cancers (Basel). 2018;10(12):487. https://doi.org/10.3390/cancers10120487.
147. Deblois G, Tonekaboni SAM, Grillo G, et al. Epigenetic switch-induced viral mimicry evasion in chemotherapy-resistant breast cancer. Cancer Discov. 2020;10:1312–29.
148. Wellenstein MD, de Visser KE. Cancer-cell-intrinsic mechanisms shaping the tumor immune landscape. Immunity. 2018;48:399–416.
149. Hargadon KM, Johnson CE, Williams CJ. Immune checkpoint blockade therapy for cancer: an overview of FDA-approved immune checkpoint inhibitors. Int Immunopharmacol. 2018;62:29–39.
150. Rohaan MW, Wilgenhof S, Haanen JBAG. Adoptive cellular therapies: the current landscape. Virchows Arch. 2019;474:449–61.
151. Kochenderfer JN, Somerville RPT, Lu T, et al. Long-duration complete remissions of diffuse large B cell lymphoma after anti-CD19 chimeric antigen receptor T cell therapy. Mol Ther. 2017;25:2245–53.
152. Chavez JC, Bachmeier C, Kharfan-Dabaja MA. CAR T-cell therapy for B-cell lymphomas: clinical trial results of available products. Ther Adv Hematol. 2019;10:2040620719841581.
153. Hartmann J, Schüßler-Lenz M, Bondanza A, Buchholz CJ. Clinical development of CAR T cells-challenges and opportunities in translating innovative treatment concepts. EMBO Mol Med. 2017;9:1183–97.
154. Ivics Z, Hackett PB, Plasterk RH, Izsvák Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell. 1997;91:501–10.
155. Wang Y, Pryputniewicz-Dobrinska D, Nagy EÉ, et al. Regulated complex assembly safeguards the fidelity of Sleeping Beauty transposition. Nucleic Acids Res. 2017;45:311–26.
156. Chicaybam L, Abdo L, Viegas M, Marques LVC, de Sousa P, Batista-Silva LR, Alves-Monteiro V, Bonecker S, Monte-Mór B, Bonamino MH. Transposon-mediated generation of CAR-T cells shows efficient anti B-cell leukemia response after ex vivo expansion. Gene Ther. 2020;27:85–95.
157. Kebriaei P, Singh H, Huls MH, et al. Phase I trials using Sleeping Beauty to generate CD19-specific CAR T cells. J Clin Invest. 2016;126:3363–76.
158. Clark KJ, Geurts AM, Bell JB, Hackett PB. Transposon vectors for gene-trap insertional mutagenesis in vertebrates. Genesis. 2004;39:225–33.
159. Converse AD, Belur LR, Gori JL, Liu G, Amaya F, Aguilar-Cordova E, Hackett PB, McIvor RS. Counterselection and co-delivery of transposon and transposase functions for Sleeping Beauty-mediated transposition in cultured mammalian cells. Biosci Rep. 2004;24:577–94.
160. Ma K, Fu D, Yu D, Cui C, Wang L, Guo Z, Mao C. Targeted delivery of in situ PCR-amplified Sleeping Beauty transposon genes to cancer cells with lipid-based nanoparticle-like protocells. Biomaterials. 2017;121:55–63.
161. Ahmad I, Mui E, Galbraith L, et al. Sleeping Beauty screen reveals Pparg activation in metastatic prostate cancer. Proc Natl Acad Sci U S A. 2016;113:8290–5.
162. Grisard E, Coan M, Cesaratto L, et al. Sleeping beauty genetic screen identifies miR-23b::BTBD7 gene interaction as crucial for colorectal cancer metastasis. EBioMedicine. 2019;46:79–93.

163. Takeda H, Rust AG, Ward JM, Yew CCK, Jenkins NA, Copeland NG. Sleeping Beauty transposon mutagenesis identifies genes that cooperate with mutant Smad4 in gastric cancer development. Proc Natl Acad Sci U S A. 2016;113:E2057–65.
164. Beckmann PJ, Larson JD, Larsson AT, et al. Sleeping beauty insertional mutagenesis reveals important genetic drivers of central nervous system embryonal tumors. Cancer Res. 2019;79:905–17.
165. Guimaraes-Young A, Feddersen CR, Dupuy AJ. Sleeping beauty mouse models of cancer: microenvironmental influences on cancer genetics. Front Oncol. 2019;9:611.
166. Chang J-H, Mou KY, Mou C-Y. Sleeping beauty transposon-mediated asparaginase gene delivery by a nanoparticle platform. Sci Rep. 2019;9:11457.

Chapter 9
Copy Number Variation in the Human Genome

Elisa Napolitano Ferreira and Caio Robledo D'Angioli Costa Quaio

9.1 Copy Number Variation: A Brief Description and Overview of its Impacts and Mechanisms of Origin

Copy number variation, or simply CNV, refers to a type of structural genomic alteration of a DNA segment detected in one individual in a number of copies different from the reference genome. CNV includes events such as duplications—that can occur in tandem or at distant sites of the genome—and deletions of one (heterozygous loss) or both alleles (homozygous loss). CNV differs from other genomic structural variants, such as inversions and balanced translocations, as it results in an unbalanced variation. The size of CNVs may vary from a few base pairs (bp), typically larger than 50 bp, to large chromosome segments and even, in extreme cases, an entire chromosome (aneuploidy) ([1, 2, 61].

From *Drosophila* to humans, CNVs are observed in different organisms, and they not only play an important role in phenotypic diversity among individuals but are also related to pathologic conditions. Given their substantial variability in size, their impact on the physiological function of an organism is also diverse and depends on many factors, for example, the size of the alteration, i.e., short segment or macroscopically detectable variant; and the genomic region in which it occurs, i.e., whether it is in a gene-rich region, a "gene desert", or a subtelomeric or centromeric region. When genes

Genes cited in this chapter: *BAZ1B* (Bromodomain Adjacent to Zinc finger domain 1B); *BRCA1* (Breast Cancer DNA repair associated 1); *CTNND2* (Catenin Delta 2); *ELN* (Elastin); *ERBB2* (Erb-B2 receptor tyrosine linase2); *GTF2I* (General Transcription Factor 2I); *GTF2IRD1* (GTF2I repeat domain containing 1); *LIMK1* (LIM domain kinase 1); *MECP2* (Methyl CpG binding protein 2); *PMP22* (Peripheral Myelin Protein 22); *UBE2QL1* (Ubiquitin-conjugating Enzyme E2 Q family-like 1).

E. N. Ferreira (✉) · C. R. D'Angioli Costa Quaio
Grupo Fleury, São Paulo, SP, Brazil
e-mail: elisa.nferreira@grupofleury.com.br; caio.quaio@grupofleury.com.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_9

are involved, the impact of the variant will depend on the genes' functions and whether a given gene was affected as a whole or in part. CNVs can affect gene dosage through gain or loss of an entire gene or multigene segments upon duplication or deletion; they can lead to gene disruption when the breakpoint of the alteration occurs within a coding gene; and they can also influence gene expression by affecting regulatory sequences of the genome, which may be close to or far away from the regulated gene [3].

Several molecular mechanisms have been described as involved in the generation of CNVs, including errors in DNA repair pathways, in recombination or in DNA replication machinery. Non-allelic homologous recombination (NAHR) is an essential mechanism of genetic diversity and can involve in the repair of DNA double-strand breaks (DSBs). Typically, NAHR occurs at multiple sites within the genome, in regions flanked by segmental duplications or low-copy repeats that are highly homologous (more than 95% identity), usually longer than 1 kb [4]. The misalignment of these highly similar sequences during meiosis or in the DSB repair process may result in the development of CNVs. Since NAHR occurs in regions containing low-copy repeats, they frequently give rise to recurrent CNVs.

On the other hand, non-homologous end joining (NHEJ) and replication-based mechanisms lead to non-recurrent CNVs of variable sizes. Essentially, human cells employ NHEJ to repair DSBs caused by ionizing radiation and reactive oxygen species. This repair mechanism comprises four steps and requires the participation of several proteins to detect the DSB, to bring the DNA molecules into physical proximity and the correct orientation, to perform the modification of the DNA ends, and to ligate the DNA segments [5]. It does not usually require any homologous segments as guides but, in some cases, small segments of 5 to 25 bp of homology are used, resulting in a mechanism known as microhomology-mediated end joining (MMEJ) [5]. Both NHEJ and MMEJ can lead to small deletions or duplications at the breakpoint junctions, leading to the development of CNVs [4].

Errors during DNA replication have also been shown to play a role in the generation of CNVs. During slippage of a single-stranded DNA at a replication fork, it may misplace such as in a self-complementary hairpin due to sequence repeats. If DNA synthesis moves forward, it may result in a deletion in the new strand; if forward and backward once the hairpin is unfold, it may lead to a duplication of the region adjacent to the repeats [4]. Serial replication slippages (SRS) can give rise to smaller complex CNVs. Additionally, rearrangements might occur between different replication forks. When a replication fork is stalled for any reason, the single-stranded lagging DNA can template-switch to another replication fork that shares at least a short homologous region. This event, called fork stalling and template switching (FoSTeS), might occur between homologous regions that are a few kilobases to several megabases in size and is believed to be responsible for the origin of many structural variants, including CNVs [6].

9.2 CNVs Are Major Contributors to Phenotypic Variation in the Human Genome

CNVs are recognized as an important source of genetic diversity in all forms of life, driving adaptive evolution in bacteria, archaea, plants and animals [7]. The first documented copy number variations were described in the early 1990s. By studying the *Drosophila Bar* gene, Sturtvant [8] suggested the occurrence of mutations that lead to different gene dosages in chromosomes as result of unequal crossing-over events. In 1936, based on cytological studies, the *Bar* mutation was confirmed as a tandem duplication mutation [9, 10].

To date, almost 100 years from the initial evidence, a huge catalogue of human CNVs has been documented, allowing the understanding of human population genetic diversity [11, 12] and human diseases.

CNVs tend to occur near low-copy repeats and are mostly located outside of genes and conserved regions of the human genome [62]. In particular, exons of genes related to diseases such as cancers are less variable than the average site within the human genome [61].

Overall, duplications are more commonly detected than deletions [11, 12, 62]. One possible explanation is that duplications have generally milder effects and are therefore better tolerated and subjected to less selective pressure [61]. Conversely, gene duplication has been described as an important mechanism of human evolution. Comparative studies across primate genomes have detected a higher number of gains than losses over evolutionary time and suggest gene duplication as a major lineage-specific gene copy number event [63].

Until recently, single-nucleotide variants (SNV) were thought to be responsible for the majority of genetic variability in humans, and CNVs, although recognized as important, were thought to be relatively rare. Nonetheless, with the evolution of technologies for studying CNVs at single-base resolution, it is now well established that CNVs are a major source of human genetic variability [64, 65].

CNVs impact more base pairs of human DNA than short-range DNA variants. On average, 1,500 common CNVs are detected in one individual, a number considerably lower than the average number of SNVs (4 to 5 millions per individual) [66–68]. However, considering that each CNV encompasses on average 20,000 bp, the extent of CNV's impact reaches at least 20 million bp (20 Mbp), in contrast to nearly 5 Mbp affected by SNP variation [11, 12, 62, 68, 69, 70].

In 2006, Redon and colleagues organized the first CNV map of the human genome based on 207 individual genomes analyzed using different array platforms. In 2015, Zarrei and colleagues published an updated version of the human CNV map by compiling high-quality published data from healthy individuals of various ethnicities. The human CNVs detected in healthy individuals as well as associated with pathologic conditions to date can be explored in public databases such the Database of Genomic Variants (DGV) (http://dgv.tcag.ca/dgv/app/home) [2]), the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER) (https://decipher.sanger.ac.uk/) [13], and the 1000 Genomes Project (https://www.internationalgenome.org/) [11, 12].

DGV is a publicly available database created in 2004 with the aim of providing a comprehensive summary of structural variation in the human genome. It presents a comprehensive catalog of human CNVs and structural variations among control individuals from populations worldwide and is continuously updated with new data from peer reviewed research studies.

DECIPHER is an interactive web-based database that aggregates both clinical and genomic information from patients in order to assist geneticists with the interpretation of genomic results, specifically those related to rare diseases and rare genomic variants. It was conceived as both a clinical and research sharing tool and compiles data from more than 36,000 cases from 270 centers.

The 1000 Genomes Project was the first high-throughput project to sequence whole genomes of a large number of individuals of different ethnicities. The project was conducted between 2008 and 2015 and sequenced 2504 individuals from 26 populations, creating the largest public catalog of human variation and genotype data to date.

These public databases have helped enormously in compiling data from different studies and helping geneticists interpret genomic data in daily practice. However, it is important to consider that the control individuals in these databases might develop a condition that appears later in life, or they might be healthy with respect to a particular condition but not be considered a "control" for a different condition, since health is dynamic. Additionally, somatic variants tend to accumulate with age and may be an additional confounding effect in some studies.

9.3 Analysis of Copy Number Variants

The initial approaches for studying CNVs were cytogenetics techniques such as fluorescence in situ hybridization (FISH) and karyotyping [71]. In [14], Jérôme Lejeune, using a cytogenetics approach, revealed that a gain of chromosome 21 was associated with Down syndrome [15]. This observation was possible because the entire chromosome 21, which represents slightly more than 48 million base pairs (48.1 Mb), is visible at microscopic resolution. Smaller chromosomal rearrangements of 5 Mb or larger may also be detected by microscopic techniques. Nevertheless, the detection of chromosomal rearrangements smaller than 5 Mb through light microscopy is a challenge; other methods, such as quantitative PCR or multiplex ligation-dependent probe amplification (MLPA), are more suitable.

By 2004, the development of new techniques such as array-CGH (comparative genomic hybridization) and SNP-array (single nucleotide polymorphism-array) and the first NGS (next-generation sequencing) platforms improved the knowledge of CNVs, since both are high-throughput technologies that allow large-scale screening of CNVs of variable sizes (Table 9.1).

Today, both NGS and array-CGH are the gold-standard technologies for CNV analysis in research and clinical settings. Although they are complex approaches that require the use of bioinformatics pipelines for the interpretation of the data, they are indicated as high-throughput screening methods for discovering CNVs associated with a specific condition. Karyotyping is also considered an appropriate screening method with simpler analysis, but it is limited to larger events. On the other hand, FISH, MLPA and qPCR are cost-effective techniques suitable for the

Table 9.1 Methods for detection of Copy Number Variations

Method	Technique	Size of CNVs detected	Throughput	Analysis
Targeted screening				
FISH	Hybridization with fluorescent probes	≥10 kb	One or a few loci	Fluorescent microscopy
qPCR	Amplification with fluorescent probes	≥50 bp to 1 kb	One or a few loci	Analysis software from the termocycler equipment
dPCR	Amplification with fluorescent probes	≥50 bp to 1 kb	One or a few loci	Analysis software from the termocycler equipment
MLPA	Probe ligation and amplification	≥50 bp	One or a few dozen loci	Coffalyser software
Genome-wide screening				
G-band karyotyping	Karyotyping	≥1 Mb	Whole genome	Microscopy
SNP-array	Hybridization	≥5 kb (*)	Whole genome	Bioinformatics pipeline
Array-CGH	Hybridization	≥1 kb (*)	Whole genome	Bioinformatics pipeline
Next generation sequencing	DNA sequencing	≥50 bp	Whole genome	Bioinformatics pipeline

(*) The size of the CNV is dependent on the resolution of the platform. The numbers represent the capacity of the high-resolution platforms

analysis of specific loci. In this section, we will describe each method in detail and comment on its advantages and limitations.

9.3.1 G-Band Karyotyping

Karyotyping is a technique for studying the number and structure of chromosomes by visual inspection, allowing the detection of abnormal numbers and/or arrangements.

The first step for karyotyping is to collect cells from the individual. Depending on the clinical setting, cells can be isolated from peripheral blood or bone marrow; amniotic fluid or chorionic villus specimens can be used for prenatal testing. These cells are then cultured *in vitro* for a short period, and by the addition of colchicine, a mitosis-inhibiting reagent, dividing cells are arrested in metaphase. Metaphase is the cell cycle phase when chromosomes assume their most condensed conformations, which facilitates visual inspection.

Cells are then disrupted, and chromosomes are fixed on a glass slide and stained to enhance visualization. Different methods may be used for staining the chromosomes. The most used to date is Giemsa staining, which produces a form known as the G-banding pattern. Giemsa incorporation into the chromosome depends on the composition of nucleotides A (adenine), T (thymine), C (cytosine) and G (guanine). AT-rich regions tend to better incorporate Giemsa and consequently result in darker-staining bands, whereas CG-rich regions appear as lighter bands. Each chromosome presents a unique G-banding pattern that facilitates chromosome identification and counting [16]. Additionally, by comparing stained chromosomes to a reference G-banding pattern from a normal karyotype, it is possible to identify certain structural variations, such as duplications and deletions of large chromosome regions, as well as translocations and inversions.

G-band karyotyping is mostly used by clinical cytogeneticists to aid in the diagnosis of specific birth defects and genetic disorders in humans by detecting large genetic changes that involve whole chromosomes or other anomalies involving several megabases of DNA.

9.3.2 Fluorescence In Situ Hybridization (FISH)

Fluorescence in situ hybridization (FISH) is another kind of cytogenetics technique that allows the visualization of structural variations of the genome. This method is based on the use of fluorescently labeled probes that bind with specific regions of DNA and thus can reveal under the microscope the occurrence of genetic amplifications, deletions and translocations [17].

Cells are obtained from blood or tissue specimens and are fixed on glass slides. They are then incubated with the fluorescently labeled probes to allow hybridization of the probes to the complementary regions of DNA. One or more probes can be used together, and repetitive sequences of DNA can be blocked to avoid nonspecific hybridization. Fluorescent signals visualized under the microscope then expose the location and quantity of the probed sequences, allowing the detection of structural variations.

For the detection of copy number variations, for instance, by counting the number of fluorescent dots specific to the centromere region of a chromosome and the number of fluorescent dots specific to a gene or locus on the same chromosome, it is possible to detect deletions, duplications and amplifications. One practical example is for subclassification of breast cancer, where the detection of amplification of the *ERBB2* gene (HER2) is indicative of a more aggressive subtype and can have an important impact on patient treatment [72].

9.3.3 Multiplex Ligation-Dependent Probe Analysis (MLPA)

The multiplex ligation-dependent probe amplification (MLPA) approach was first described in 2002 by Schouten and collaborators [18]. The method is based on hybridization of DNA probes, followed by amplification and fragment analysis by capillary electrophoresis.

For each assayed region, two probes are designed, targeting adjacent segments of DNA. Upon hybridization, the probes are positioned adjacent to one another, allowing their ligation into one unique probe. This step gives high specificity to the method since the ligation will only occur once both probes are accurately aligned with the genome. Next, the unique longer probe is linearly amplified by universal primers, and the generated fragments are analyzed by capillary electrophoresis.

The universal primers are complementary to sequences at the ends of the probes. The use of universal primers allows different pairs of probes to be used and amplified in parallel in a single MLPA experiment. Additionally, each pair of probes must produce a fragment of a different size to allow the amplicons to be distinguished by electrophoresis. The amount of each amplified fragment correlates to the DNA dosage in the sample tested. Therefore, by quantifying these fragments, it is possible to detect regions of copy number duplications and deletions.

Several commercial MLPA assays are available through a company named MRC Holland for the study of numerous human diseases. MLPA is considered a high-confidence method for CNV analysis of specific regions, with high resolution for small CNV events.

9.3.4 Polymerase Chain Reaction (PCR)

Polymerase chain reaction (PCR) is a flexible, simple, low-cost approach that is widely used in multiple molecular studies. PCR consists of recreating the conditions of DNA replication *in vitro* to amplify a DNA fragment of interest by producing multiple copies. This method was originally introduced by Saiki in [19] and automated by Mullis in [20] and is considered one of the most basic and important tools in molecular biology.

From a small amount of DNA, a specific region of the genome is amplified *in vitro* by using specific synthetic oligonucleotides, a thermostable polymerase enzyme (usually *Taq* DNA Polymerase), dNTPs (deoxyribonucleotides A, T, C and G) and buffers, salts and co-factors to enhance the polymerase reaction. The oligonucleotides are complementary to the region of interest and serve as the initiators of DNA synthesis *in vitro*. *Taq* DNA polymerase is purified from the bacterium *Thermus aquaticus*, which is a microorganism that lives in habitats with extremely high temperatures. The amplification occurs in multiple cycles, where at each cycle, the number of DNA molecules is doubled, leading to an exponential amplification of the molecules.

The PCR cycles consist of a DNA denaturation step by heating at approximately 95 °C, the hybridization of the oligonucleotides to the complementary regions of the DNA at a lower annealing temperature (the annealing temperature depends on the base composition of the oligonucleotide and usually ranges from 50 to 70 °C) and, finally, an extension step where the actual polymerase activity of incorporation of dNTPs into the growing chain of DNA occurs at temperatures between 68 to 72 °C.

Conventional PCR, as described above, can be helpful in detecting structural variations by amplifying the breakpoints of the events. For CNV detection, quantitative PCR and, more recently, digital PCR are more appropriate. Quantitative PCR uses a set of fluorescently labeled probes that generates a quantifiable signal at the end of each amplification cycle and allows the relative quantification of a target and a reference fragment. By using a DNA region with a known copy number (normally diploid) as a reference, the concentrations of the target and reference gene can be used to estimate the copy number of the target [21].

Digital PCR is similar to quantitative PCR; however, prior to amplification, the reaction mix is partitioned into several thousand reactions (emulsified droplets or physical compartments) to dilute the target DNA such that each partition contains zero or one copy of the target. After amplification, every partition is investigated to determine whether it is positive or negative for both the target and the reference region, resulting in a digital counting of a binary outcome (positive or negative). Digital PCR is not a screening method; however, it gives a highly precise, absolute quantification of DNA copy number in a fast, easy and low-cost approach [22].

9.3.5 Hybridization-Based Microarray Approaches

Hybridization platforms consist of DNA fragments of known sequences that are immobilized to a solid platform in a known position and order, creating a microarray. DNA from a specific sample of interest is fragmented and labeled, and these labeled DNA molecules are incubated with the array under specific conditions (buffers and temperature). By complementary hybridization, the labeled DNA is attached to the solid platform, and the fluorescent signals and intensities are retrieved with a scanner. The amount of DNA in the sample can be inferred by the intensity of the signals obtained in each specific spot on the array platform.

In a comparative genomic hybridization (CGH) array, DNA from two samples is used. The DNA from a reference sample is labeled with one fluorophore, and DNA from the test sample is labeled with a different fluorophore. Both types of DNA are incubated with a single microarray for competitive hybridization. After scanning, the signal from the test sample is compared to that of the reference sample, and the ratio indicates the copy number.

At copy number duplications, a higher intensity of the test sample compared to the reference is detected. Conversely, when a higher intensity of the reference sample is detected, it indicates a deletion on the test sample. Regions of equal intensity represent neutral copy number.

Historically, CGH arrays were composed of bacterial artificial chromosome (BAC) clones, which are large segments of DNA (≥100 kb) and therefore produce only low-resolution analysis of CNVs [62, 64]. More recently, long oligonucleotides (approximately 50–120 bp long) have been used instead, offering higher resolution; these arrays are able to detect CNVs as short as 500 bp [69]. With advances in array production technology, such as the development of high-density probe arrays and the possibility of customizing platforms, this technology is currently one of the preferred methods for investigating copy-number alterations among children with development delay syndromes in clinical practice.

In contrast to CGH, SNP microarray platforms interrogate specific SNPs using one sample at a time. Initially, SNP arrays were applied for genotyping, population genetics and epidemiology studies, but later, their use was broadened to include the detection of copy number changes [23]. SNP arrays contain allele-specific oligonucleotide probes immobilized onto their surface. For each locus of interest, two probes are designed, one probe specific to one of the alleles and the other specific to the alternative allele, allowing the detection of homozygous and heterozygous genotypes. Using a specific bioinformatics pipeline, copy number variation can be accurately predicted from SNP array platforms. The results tend to have less noise than the CGH array results; however, the precise breakpoint of the alteration is harder to define. Additionally, it is possible to combine CGH and SNP arrays in a single platform to enhance the accuracy of CNV detection.

9.3.6 Next-Generation Sequencing (NGS)

The advent of next-generation sequencing (NGS) technologies has revolutionized the analysis of human genome alterations, given that they allow the study of a multitude of alteration types, including CNVs. NGS approaches rely on high-performance DNA sequencers permitting the simultaneous sequencing of an extraordinarily large number of molecules. A great innovation of these technologies is that they dispense the bacterial cloning step that was required for genome sequencing based on the Sanger sequencing method (see Chap. 3). Instead, the amplification of the DNA molecules was replaced by emulsion PCR or PCR of templates conjugated to a solid matrix (bridge-PCR). The sequencing approach varies, and depending on the platform, fluorescent labeling, pH variation or changes to an electrical current may be used for sequence detection.

The large amount of data retrieved from NGS requires sophisticated bioinformatics pipelines for analysis [24]. Briefly, each read generated by sequencing will be evaluated for quality parameters, and then mapped to the reference human genome sequence. By comparing the composition of the bases sequenced to the reference genome and the number of reads mapping to each genomic coordinate position (sequencing depth), it is possible to identify molecular alterations. By analyzing the read depths of different genome regions, it is possible to detect regions with significantly higher or lower than expected sequencing depths, revealing copy number expansion and deletion, respectively.

Additionally, analyzing the alignment of the reads to the reference genome can detect fragments with split alignments, suggesting the deletion of the skipped region, or with duplicated stretches of DNA, suggesting copy number gains. Furthermore, the development of paired-end sequencing, which allows the sequencing of DNA fragments from both the 3′ and 5′ ends, has enhanced the detection of CNVs and other structural variants [25]. In this approach, the deletion of a genomic region would result in the alignment of the paired reads at regions closer than expected, and conversely, a duplication/insertion would cause the mapping of the pair of reads at a greater distance than expected. Different algorithms focused on the detection of CNVs have been reported, and the use of more than one algorithm in combination is an interesting approach.

Although complex and expensive, NGS platforms have the advantage of allowing the comprehensive evaluation of a patient's genome through the detection of several types of alterations in a single experiment, including point mutations, small insertions/deletions (indels), rearrangements and copy number variants.

9.4 CNVs and Human Diseases

The first studies of submicroscopic chromosomal rearrangements demonstrated that CNVs are frequent in all individuals, are spread throughout the genome (though they may occur more frequently in some chromosome segments), may or may not be inherited (if not inherited, they are called *de novo* CNVs), contribute to human genetic diversity and may be associated with human diseases. Several human diseases are now known to be associated with CNVs, including diseases with classical Mendelian inheritance (e.g., autosomal dominant, autosomal recessive and X-linked) and complex inheritance, such as autism spectrum disorders (ASD) and mental illnesses.

The advancement of CNV studies and their increased application in clinical diagnostics have led to a great expansion of the number of genetic diseases that have been directly associated with CNVs [26]. The impact of a CNV in human disease depends on the number of genes it encompasses, the effects that these genes exert and the molecular mechanism of the rearrangement (gain or loss of genetic material). Many genes are dosage sensitive, and an alteration in their copy number results in altered expression of the corresponding gene product. This is probably the most common molecular mechanism underlying CNV-mediated pathogenesis. The altered expression of several dosage-sensitive genes has been associated with human diseases, and several examples are discussed below. Other atypical mechanisms will also be reviewed.

Several human diseases caused by pathogenic CNVs are cataloged in OMIM [27], which is an online compendium of human genes and genetic phenotypes (10). OMIM is one of the main knowledge sources that geneticists reference to understand human Mendelian diseases, their molecular mechanisms and their implications for health. Its use is especially valuable for the clinical follow-up of an individual with a genetic disease. The OMIM number (OMIM#*number*) is given for several of the diseases discussed in this chapter to facilitate further study on the OMIM platform, which can be freely accessed at www.omim.org. The authors recommend that readers explore the OMIM platform and its connections with other databases, especially GeneReviews (an online, up-to-date point-of-care resource for clinicians covering molecular diagnosis, management and genetic counseling).

9.4.1 CNVs and Neurological Diseases: The PMP22 *Gene as an Example*

A neurological disease that involves peripheral nerves, Charcot-Marie-Tooth disease type 1A (CMT1A), is a good example of a human disease associated with a specific CNV with classical autosomal dominant inheritance (OMIM#118220). Individuals from families with CMT1A present progressive degeneration of the peripheral nerves that cause sensory loss and weakness of more distal parts of limbs. These patients usually present symptoms before the second decade of life, frequently beginning with weakness and atrophy in the muscles of the hands and feet. Deformity of the feet (*pes cavus*), foot drop, decreased sensation in and numbness of fingers and toes are very common early presentations [28]. Although the disease progresses throughout the patient's life, very few patients become wheelchair dependent, and the life span is usually normal.

Molecular studies of families with CMT1A made tremendous advancements in the 1990s [29], when researchers observed that this condition was associated with a small CNV on the short arm of chromosome 17 (17p12). This CNV was found to be a gain (e.g., a microduplication) of approximately 1.5 Mb that involves an important gene, *PMP22*. This gene encodes a structural protein component of myelin that is important for the maintenance of myelinated fibers in the peripheral nervous system. A microduplication involving the *PMP22* gene, such as those harbored by patients with CMT1A, leads to overexpression of this gene product and consequently increased levels of PMP22 protein, which disrupts the myelin fibers produced by Schwann cells. Therefore, an increased dosage of the *PMP22* gene product underlies the etiology of CMT1A.

Interestingly, decreased dosages of this same gene, which may be found in individuals harboring microdeletions of chromosomal region 17p12, also have an impact on human health. This microdeletion leads to underexpression of this gene product and consequently decreased levels of the PMP22 protein, which also disrupts the function of myelin fibers, making nerves susceptible to conduction block when they are compressed.

These patients manifest a different form of autosomal dominant neurological disease of peripheral nerves, called hereditary neuropathy with liability to pressure palsies (HNPP) (OMIM#162500). This condition was first described in 1947 in a three-generation family in which individuals had recurrent neuropathy of the peroneal nerve, manifested as foot drop and sensory alteration of the foot or the outer part of the upper or lower leg, after spending time in a kneeling position digging potatoes [30]. Indeed, HNPP typically leads to attacks of numbness and muscular weakness triggered by minor compression on the affected nerve, such as prolonged positioning of the limb. The most vulnerable nerves are the peroneal and ulnar nerves [31].

Therefore, the *PMP22* gene is sensitive to both increased and decreased dosages. The terms haploinsufficiency and triplosensitivity describe cases in which the loss of one copy of a gene or the presence of an additional copy of a gene, respectively, causes a phenotype. In summary, individuals with microduplications of 17p12 have three copies of *PMP22* and consequently an overproduction of PMP22 protein, whereas patients with microdeletions of 17p12 have only a single copy of this gene and diminished production of the protein. Although microduplications and microdeletions represent opposite mechanisms, both CNVs alter *PMP22* function and have an impact on human health.

9.4.2 CNVs and Cancer

Cancer describes a group of different diseases characterized by uncontrolled cellular proliferation that may lead cells to invade other tissues and organs, causing dysfunction, inflammation and occasionally leading to death [32, 33]. Approximately 5% to 10% of all cancer cases are hereditary. Individuals who inherit genetic alterations associated with the onset of cancer (e.g., familial cancer syndrome) present a considerably higher risk of developing cancer throughout life compared to the general population.

Hereditary breast and ovarian cancer (HBOC) is a good example of an autosomal dominant syndrome predisposing to cancer (OMIM#604370). Several genes are associated with this form of familial cancer. The *BRCA1* gene has an important role in several families because alterations in this gene are relatively common, and its impact on life expectancy and health is substantial. While the risk of breast cancer throughout life is approximately 5% to 11% for women on average, women with a pathogenic alteration in *BRCA1* have a lifetime cancer risk of 85% [34]. Additionally, these women also have a 45% lifetime risk of developing ovarian cancer. The identification of pathogenic alterations in HBOC-associated genes are important for genetic counseling and to inform the patient on the indication, risks and benefits of prophylactic surgery (oophorectomy and mastectomy) before the onset of cancer, as well as to implement clinical and psychological surveillance.

The *BRCA1* gene encodes the BRCA1 protein, which has important interactions with several other proteins involved in cellular pathways. Loss of function of *BRCA1* (e.g., haploinsufficiency) results in defects in DNA repair, transcription, defective cell-cycle regulation, chromosome damage and, ultimately, an increased risk of developing certain types of cancer [35]. Interestingly, gain of function of *BRCA1* does not have any known impact on human health. Therefore, the *BRCA1* gene is considered haploinsufficient but not triplosensitive.

Although the large majority of alterations leading to loss of function in *BRCA1* correspond to sequence variants, pathogenic CNV is also an important mechanism in several families, especially those of Latin American and Caribbean descent [36]. Whole-gene deletion of *BRCA1* has been reported previously. We present below two atypical mechanisms in which a CNV can cause human disease through effects on *BRCA1*.

Small deletions of one or more exons within *BRCA1,* called *BRCA1* intragenic deletions, are important disease mechanisms in several families with hereditary breast cancer. Intragenic deletions may inactivate the gene, having the same effect as other null sequence variants of that gene. Intragenic deletions in *BRCA1* correspond to almost 10% of all alterations in this gene [36]. Considering that these rearrangements cannot be studied by routine sequencing techniques, the inclusion of assays for the comprehensive detection of CNVs is highly recommended by experts for the study of HBOC. In this *BRCA1* example, we observe how an intragenic CNV may have the same effect as a frameshifting variant, effectively inactivating the gene and leading to haploinsufficiency.

A gene promoter is a DNA sequence to which the RNA polymerase complex binds and then initiates transcription of the DNA sequence downstream synthesizing RNA (see Chap. 4). Genetic alterations, such as CNVs, within the gene promoter region could potentially disrupt the normal expression of the respective encoded protein. Although uncommon, deleterious CNVs within the promoter region of *BRCA1* have an important impact on several HBOC families with European ancestry [37, 38]. Considering that the clinical impact of promoter disruption is very similar to those of a pathogenic sequence variant or even whole-gene deletion, CNV studies of the promoter region of *BRCA1* are routinely performed in laboratories around the world. *BRCA1* promoter microdeletion is an example of how CNVs may also alter the expression of genes that are upstream or downstream of the CNV by disrupting regulatory elements such as enhancers or promoters.

9.4.3 CNVs and Malformative Syndromes

Congenital birth defects (CBD) occur in approximately 3% of newborns and may result in varying degrees of physical, mental or developmental disabilities [39]. As severe defects may be fatal early in life, this group of anomalies has become one of the most frequent causes of death in childhood in developed countries. CBD can cause lifelong disability, which may have a significant impact not only on the affected individuals but also on their families, society and healthcare systems. Because children with CBD often require a variety of health services, including medical care, rehabilitation therapies, special medications and so on, several countries have created special regulations and policies for the diagnosis, coordinated care, counseling and support for CBD patients and their families. The importance of this subject has been recognized by the World Health Organization, which has developed normative tools, guidelines and a global plan of action [40].

CBD may result from a variety of events, including genetic alterations and nongenetic exposures during pregnancy, such as exposure to medications (e.g., thalidomide) or chemicals (e.g., maternal use of alcohol) or maternal infection (e.g., Zika virus, mononucleosis). Although nongenetic exposure is an important subject for health, it is not related to genetic origin and will not be discussed in this chapter. Our discussion will be limited to congenital malformations of genetic origin, which are deleterious physical anomalies attributable to altered embryonic or fetal development (such as cell differentiation, migration or other important regulation) that are manifested at birth. A combination of malformations involving one or more body parts in a recognizable pattern is referred to as a malformative syndrome.

The genetic basis of malformative syndromes also varies; it includes large chromosomal rearrangements, deleterious sequence variants in genes important for embryonic development and pathogenic CNVs. CNVs, both deletions and duplications, have long been recognized as underlying etiologies in well-characterized genetic disorders and syndromes [26].

Williams-Beuren syndrome (WBS) (OMIM#194050) is one of the most frequent multisystem disorders associated with a pathogenic CNV. Although it has been known since 1961 as a syndrome characterized by supravalvular aortic stenosis, intellectual disability, and distinctive facial features [41], it was only in 1993 that it was first discovered that WBS was caused by the haploinsufficiency of genes within a microdeletion at 7q11.23 [42]. This chromosomal region is prone to recurrent chromosomal rearrangements, including the microdeletion that causes WBS.

Patients with WBS present several specific characteristics of varying degrees of severity. The facial features (frequently referred to in genetics as *facies*) appear to be "elfin", with a broad forehead, short nose, full cheeks, stellate iris pattern, flat nasal bridge, malar flattening, full lips, long and smooth philtrum, pointed chin and wide mouth [43]. WBS patients usually demonstrate an extroverted personality in infancy and are often described as having a happy, outgoing personality and proneness to interact readily with strangers, but this may disappear later in life. Another remarkable feature in WBS is cardiovascular involvement, characterized by the presence of stenotic arteriopathy. This phenomenon is very specific to WBS and is found in approximately 80% of patients, with supravalvular aortic stenosis being the most common form, followed by renal artery stenosis. Other multisystem involvements that these patients may present, along with their respective frequencies, are as follows: intellectual disability (75%), failure to thrive (75%), sleep problems (65%), strabismus (50%), urinary anomalies (50%), lax joints, anomaly of calcium metabolism, and diabetes mellitus, among others [44].

The recurrent deletion within the 7q11.23 region in patients with WBS comprises either 1.55 Mb, found in 90% of individuals, or 1.84 Mb, found in the remaining 10%. There may be other atypical, less common deletions, and phenotypes may vary depending on the extent of the microdeletion: a more severe phenotype with lower cognitive ability is observed in individuals with very large deletions

(>2–4 Mb), while those with partial deletions may not have intellectual disability [45, 46].

The study of the genes located in 7q11.23 elucidated the molecular mechanisms responsible for the clinical features and deeper genotype-phenotype correlations in WBS. This region comprises 25 important genes, some of which are responsible for specific clinical features. The *ELN* gene encodes elastin, an elastic protein that is highly expressed in the cardiovascular system, especially in the great arteries. Several studies have demonstrated that haploinsufficiency of *ELN* underlies the cardiovascular involvement in WBS, often referred to as elastin arteriopathy [47]. In this context, haploinsufficiency of *ELN* leads to lower production of elastin in the connective tissue of the great arteries and subsequent stenotic arteriopathy of the aorta, renal arteries and, less commonly, other arteries.

The *ELN* gene in WBS is the critical region for arteriopathy. "Critical region" is the term commonly used when the involvement of a gene or a group of genes is specifically associated with a clinical impact. Critical regions represent critical dosage-sensitive elements of the genome that may be responsible for some of the deleterious phenotypes observed for pathogenic CNVs. Other critical regions for specific characteristics in WBS are as follows: *GTF2I*, responsible for intellectual disability; *LIMK1*, for abnormality of visuospatial constructive cognition; *BAZ1B*, for anomalies of calcium metabolism; and *GTF2IRD1*, for typical facial characteristics [44].

As we can see in WBS, the clinical impact of the microdeletion depends upon the number and the identities of the genes involved. In this group of patients, those without involvement of the *ELN* gene will likely not develop cardiovascular anomalies, while those without *GTF2I* involvement are more likely to have near normal intelligence. On the other hand, individuals with haploinsufficiency of *GTF2IRD1* alone are thought to present facial characteristics of WBS without other prominent multisystem involvement [48].

Cri-du-chat syndrome (OMIM#123450) is a clinical condition associated with a CNV on the short arm of chromosome 5. It was first described in 1963 by the French scientist Lejeune and his team [49] to include signs such as diminished head circumference (microcephaly), round face, ocular hypertelorism, small jaw, epicanthal folds, hypotonia and severe intellectual disability. This genetic syndrome was named after the high-pitched cry, which is one of the most typical characteristics found in newborns with this syndrome and resembles the cry of cats.

Several patients have been described in the literature since the first publication presenting varying degrees of clinical involvement, but the great majority had the unique cat-like cry. The deletions can vary in size from extremely small, involving only band 5p15.2, to the entire short arm [50]. Using different methodologies, the critical region for the cat-like cry was mapped to a specific candidate gene, *FLJ25076*, which encodes an ubiquitin-conjugating enzyme (UBE2QL1) involved in protein degradation [51].

The critical region for severe intellectual disability was attributed to the *CTNND2* gene, which encodes a neuron-specific protein expressed early in development and involved in cell motility [50].

9.4.4 CNVs and Neurodevelopmental Disorders

Autism spectrum disorder (ASD) refers to a group of developmental disorders characterized by difficulties with social interaction, impairments of communication and engagement in repetitive, compulsive or ritualistic behaviors [52, 53]. ASD affects 1% to 2% of children globally and presents a great challenge for families, society and health policy. Symptoms in children with ASD arise early in the first years of life, generally manifesting as less attention to social stimuli, delayed or even absent smile, little interest in looking at others and difficulties in responding to their own name. Severity may vary from person to person, but almost one-third of these individuals will not develop enough speech to meet communication needs for independence, and their lifelong disability may require substantial social and educational support [53].

It is known that ASD has a strong genetic basis, although complex and still unknown for most patients [52]. This complexity may be partially explained by interactions among multiple genes, the environment and epigenetic factors that influence gene expression. Therefore, ASD corresponds to a group of conditions with different genetic backgrounds and inheritance patterns: while some cases present with classical Mendelian inheritance and are caused by single-gene mutation or pathogenic CNV, others may derive from chromosome abnormalities, oligogenic interactions or multifactorial inheritance; and yet many cases may arise exclusively from environmental exposure (such as exposure during pregnancy to infection, alcohol or illicit drugs).

Cytogenetic abnormalities visible through light microscopy can be found in up to 7% of children with ASD [54], while submicroscopic CNVs are the underlying etiology in almost 15% of these patients [55]. The majority of CNVs associated with ASD occur *de novo* (i.e., are not inherited from parents), but there are some CNVs that may be inherited in either autosomal dominant, X-linked or even autosomal recessive patterns. Considering the increased power of CNV analysis for detecting underlying etiologies in ASD patients, since it can detect not only most alterations observed by classical cytogenetics but also submicroscopic alterations, it has been recommended as a first-tier testing method in expert guidelines [56]. The etiology determination of ASD may improve patient care, clinical management and genetic counseling for the family.

More than 100 different CNVs have been associated with ASD. Several of these alterations occur in "ASD hot spots", which are regions containing important neurodevelopmental genes, more prone to undergo rearrangements. Some CNVs may determine additional manifestations, such as other neurological symptoms (e.g., seizures), intellectual disability, dysmorphisms, growth anomalies, organ malformations, and metabolic alterations. The mechanism by which CNVs determine ASD varies widely. Two mechanisms will be discussed more deeply: dosage sensitivity, when a CNV involves a dosage-sensitive gene, and participation of CNVs in recessive disorders.

A good example of dosage sensitivity in ASD is the *MECP2* gene, which is located on the long arm of the X-chromosome (Xq28), is crucial for the development of neural circuitry at embryonic stages and sensitive to copy number gains or losses. Most alterations involving *MECP2* occur *de novo*, although a small percentage is inherited maternally.

Hemizygous loss of *MECP2* (such as a microdeletion or pathogenic sequence variant involving the single copy of this gene in males) is associated with a very severe and often lethal form of encephalopathy. Heterozygous losses (microdeletions or pathogenic sequence variants involving one copy of this gene in females) are associated with Rett syndrome (OMIM#312750), which is one of the most common causes of ASD in females and is generally associated with other symptoms, including microcephaly, epilepsy, short stature, and unsteady gait, among others.

Copy-number gain of *MECP2* in males and females is also associated with neurodevelopmental anomalies. Patients harboring a microduplication or even a microtriplication of this gene present Lubs X-linked mental retardation syndrome (OMIM#300260), which generally manifests with hypotonia, feeding difficulties, gastroesophageal reflux, constipation, severe intellectual disability, and recurrent respiratory infections. Lubs syndrome presents incomplete penetrance in females and complete penetrance in males.

Several autosomal recessive metabolic abnormalities have been reported in ASD, generally presenting early in life and associated with other neurological symptoms, such as seizures, neurodegeneration, parkinsonism, and failure to thrive. CNV analysis plays an important role in the diagnosis of such conditions because it may reveal X-linked or recessive disorders, wherein a deletion of one allele of the gene unmasks a point mutation on the other allele, resulting in the disease. Although rare, CNVs involving genes associated with metabolic diseases have been described [57]. This subject is particularly important because several metabolic diseases can be prevented with proper treatment, and early diagnosis and treatment may prevent further neurological deterioration.

In summary, we have studied several examples of the relationship between CNVs and human diseases and their main molecular mechanisms of pathogenicity. Alteration in the copy number in dosage-sensitive genes results in altered expression of the corresponding gene product. Intragenic CNV may have the same effect as a disruptive point mutation, effectively inactivating the gene and leading to haploinsufficiency. CNVs may also alter the expression of genes that are upstream or downstream of the CNV by disrupting regulatory elements such as enhancers or promoters. Finally, CNVs may reveal recessive disorders, wherein a deletion of one allele of the gene unmasks a point mutation on the other allele, resulting in disease.

9.5 Clinical Application of CNV Study in Human Disease

In this section, we will discuss three practical examples of CNV study in the clinical setting, exploring the process from the evaluation of the patients by clinicians to the ordering of genetic testing and then to the laboratory interpretation of CNV results.

9.5.1 Clinical Case 1

A three-year-old boy was referred to a genetics clinic for evaluation of neurodevelopmental delay and some signs of ASD. Clinicians use the term neurodevelopmental delay when a child presents a delay in achieving certain development milestones. In this case, the boy started to walk at 1 y 6 mo and said his first words at 2 y, while both of these milestones are expected around the age of 1 y. Except for an increased height for his age and mildly decreased muscle tone (hypotonia), his physical examination was unremarkable: his face had no signs of dysmorphism, and extremities, genitals, skin, joints, heart, abdomen and spine were all normal.

His first set of exams included an echocardiogram to properly analyze the heart, brain MRI for the presence of abnormal superior structures, spine radiography, abdominal ultrasonography and thorough metabolic workup including thyroid function tests, liver enzymes, complete blood count, glucose, lactate, urinalysis, and muscle markers. These sets of tests are recommended for all individuals with neurodevelopmental disorders to investigate the presence of malformations in internal organs and any metabolic alterations that might aid in the diagnosis. This patient had normal results for all exams requested.

Following experts' guidelines, the clinician ordered a chromosomal microarray for copy number analysis across the genome. This approach is routinely recommended as a first-tier clinical practice for the evaluation of disease-causing alterations in patients with developmental delay, ASD, intellectual disability or congenital anomalies [58].

Chromosomal microarray was performed using the Agilent CGH + SNP Microarray 400 K platform. This platform usually calls approximately 60 CNVs per sample. Most of them are known to be benign and are considered common polymorphisms (i.e., they occur in more than 1% of the population). One very useful and continuously updated tool to investigate the frequency of a specific CNV in general populations is the Database of Genomic Variants (DGV), which provides a catalog of control data for studies and analyses aiming to correlate genomic variation with phenotypic data [2]. Although this tool can be accessed through the website (http://dgv.tcag.ca/dgv/app/home), DGV has already been integrated into several genome browsers, such as the UCSC Genome Browser, which facilitates the exclusion of CNVs found at high frequency in controls, keeping the focus on those rare CNVs.

In this specific case, only one CNV was absent among controls from DGV. This rare CNV was a full duplication of chromosome X, spanning from chromosomal positions 1 to 150,858,234 in the hg19 version of the genome, representing more than 150 Mb (Fig. 9.1a). In other words, this patient presented a whole extra copy of chromosome X, which is the characteristic finding of Klinefelter syndrome. This aneuploidy could have been detected by a classic microscopic cytogenetic analysis (e.g., G-band karyotype), which would also have detected the extra X chromosome; the karyotype would have been 47,XXY instead of the normal 46,XY for males.

Klinefelter syndrome is one of the classical sex chromosome syndromes in humans and is clinically characterized by varying degrees of tall stature, infertility, gynecomastia, hypotonia and anomalies of neurodevelopment (see Chap. 2) [59]. Though relatively common, Klinefelter syndrome is not cataloged in OMIM because it is not an example of Mendelian inheritance.

9.5.2 Clinical Case 2

A three-month-old girl was referred for investigation of multiple congenital anomalies, dysmorphisms and neurodevelopmental delay. Her medical issues arose prenatally, when obstetric ultrasound revealed abnormal growth, heart malformation (ventricular septal defect), brain malformation with enlarged ventricles and agenesis of the corpus callosum and deformity of the feet (club feet). After birth, she required multidisciplinary care in the neonatal intensive care unit and received emergency surgical correction of her heart defect in her first month of life.

Her physical examination indicated several facial characteristics that differed from those of her parents and standards expected for her age, overall suggestive of syndromic origin. These abnormalities of facial features are often referred to as facial dysmorphisms. Her dysmorphisms included diminished head circumference (microcephaly), asymmetric ears, deep-set eyes, straight eyebrows and pointed chin.

Chromosomal microarray was performed using the Agilent CGH + SNP Microarray 400 K platform. Again, several CNVs were found, and the first step consisted of eliminating those known to be benign. Some CNVs are not polymorphic (<1% frequency in controls) but are considered benign. Other important characteristics [58] are often considered to classify a CNV as benign, such as the following:

1. the CNV does not span protein-coding genes or any known functionally important elements;
2. the CNV overlaps completely or partially with established benign genes or genomic regions curated by experts, independently of the frequency of the CNV in controls;
3. for inherited CNVs, if the CNV does not segregate with the disease in one or several families, there are two possible scenarios that point to benign impact:
   (a) the CNV is found in an affected proband, but not in another individual in the proband's family who is also affected with a consistent, specific, well-defined phenotype;
   (b) the CNV is found in the affected proband and is also found in another individual in the proband's family who is unaffected with the specific, well-defined phenotype observed in the proband;
4. Population studies show no statistically significant difference in the frequency of the CNV between cases and controls.

After eliminating irrelevant CNVs, a microdeletion in the short arm of chromosome 1, located at subtelomeric band 1p36 and spanning from chromosome positions 1 to 4,310,992, according to genome version hg19, remained (Fig. 9.1b). In total, this microdeletion segment spanned slightly more than 4.3 Mb, which would not be visible by standard microscopic cytogenetics analysis.

As a 4.3-Mb microdeletion was identified in 1p36, it was advisable to explore details of this region in a genome browser, such as the UCSC Genome Browser (Fig. 9.1b). Deeper analysis of this region demonstrated that it included a considerably high number of important genes. The next question was: could a microdeletion within this region alter the function of any gene? To answer that question, we investigated whether the region included any genes sensitive to haploinsufficiency.

The Clinical Genome Resource (ClinGen) consortium has generated the ClinGen Dosage Sensitivity Map by curating genes and regions of the genome to assess whether there is evidence to support the dosage sensitivity of these genes/regions. This tool is also integrated into several genome browsers and can be accessed through the website dosage.clinicalgenome.org. When using ClinGen to explore the microdeletion region found in this patient, we observed that the 1p36 region is clearly sensitive to haploinsufficiency and that microdeletions involving it are associated with a known contiguous gene deletion syndrome: chromosome 1p36 deletion syndrome (OMIM#607872). Chromosome 1p36 deletion syndrome is a widely known syndrome associated with multiple congenital anomalies and intellectual disability [60].

When a CNV completely overlaps with a dosage sensitivity region, whether a microdeletion completely overlaps with an established haploinsufficiency gene/genomic region or a microduplication overlaps with an established triplosensitivity gene/genomic region, it classifies as disease-causing or pathogenic [58]. As the microdeletion identified completely overlaps with a region sensitive to haploinsufficiency associated with a known human disease, we could conclude that the clinical findings of our patient were due to the 4.3-Mb deletion, and the final diagnosis was chromosome 1p36 deletion syndrome.

9.5.3 Clinical Case 3

Clinical information, although critically important for proper genotype-phenotype correlations when analyzing genetic tests, is not always provided to diagnostic laboratories. The lack of clinical information increases the challenges of CNV analysis. In this case of a four-year-old girl, the only clinical information provided was "multiple congenital malformations", and therefore the analysis was focused on CNVs associated with a malformative syndrome.

Again, we performed a chromosomal microarray using the Agilent CGH + SNP Microarray 400 K platform and eliminated those CNVs found to be benign from the analysis. After these initial steps, we found a relevant CNV not present in the control databases: a gain of genomic material at the end of the long arm of chromosome 3, spanning approximately 22.8 Mb from chromosomal positions 175,175,935 to 199,033,185 (Fig. 9.1c). However, instead of a duplication (e.g., three copies) of

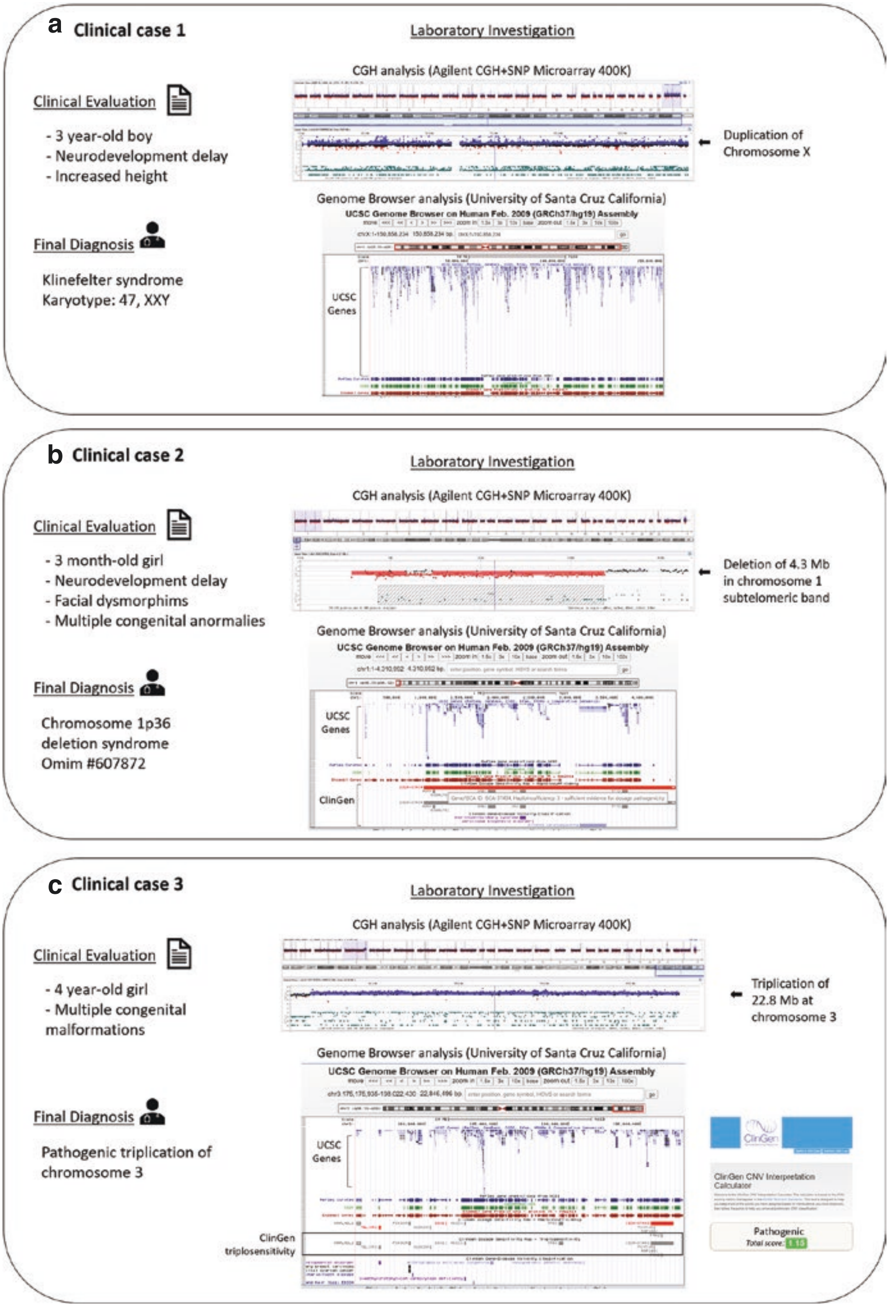

a Clinical case 1
Laboratory Investigation
CGH analysis (Agilent CGH+SNP Microarray 400K)
Clinical Evaluation
- 3 year-old boy
- Neurodevelopment delay
- Increased height
Duplication of Chromosome X
Genome Browser analysis (University of Santa Cruz California)
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
Final Diagnosis
Klinefelter syndrome
Karyotype: 47, XXY
UCSC Genes
b Clinical case 2
Laboratory Investigation
CGH analysis (Agilent CGH+SNP Microarray 400K)
Clinical Evaluation
- 3 month-old girl
- Neurodevelopment delay
- Facial dysmorphims
- Multiple congenital anormalies
Deletion of 4.3 Mb in chromosome 1 subtelomeric band
Genome Browser analysis (University of Santa Cruz California)
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
Final Diagnosis
Chromosome 1p36 deletion syndrome
Omim #607872
UCSC Genes
ClinGen
c Clinical case 3
Laboratory Investigation
CGH analysis (Agilent CGH+SNP Microarray 400K)
Clinical Evaluation
- 4 year-old girl
- Multiple congenital malformations
Triplication of 22.8 Mb at chromosome 3
Genome Browser analysis (University of Santa Cruz California)
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
Final Diagnosis
Pathogenic triplication of chromosome 3
UCSC Genes
ClinGen triplosensitivity
ClinGen
ClinGen CNV Interpretation Calculator
Pathogenic
Total score: 1.15

Fig. 9.1 Examples of three clinical cases. For each patient we show the description of the major clinical characteristics, the laboratory results and the final diagnosis. For all cases, a CNV was identified using CGH + SNP Microarray platform (Agilent Technologies). The genes involved in the CNV were investigated using the Genome Browser from Santa Cruz California University. The Clingen calculator was also used in the third case. (**a**) Clinical case 1. (**b**) Clinical case 2. (**c**) Clinical case 3

this chromosomal region, our patient presented a triplication; in other words, our patient presented four copies of that segment—a total gain of 45.6 Mb (2 x 22.8 Mb) (Fig. 9.1c).

This chromosomal region is rich in regulatory elements and genes. Exploring the region with the UCSC Genome Browser revealed that it contains 359 genes curated by RefSeq. Therefore, this patient presented a triplication of genomic material that was remarkable not only in the extent of the genomic regions but also in its high density of important coding elements. One possible question raised is as follows: is there any triplo- or tetrasensitive element? The ClinGen Dosage Sensitivity Map may help to answer this question since, as already discussed in Case 2, when a microduplication overlaps an established triplosensitivity gene/genomic region, this situation is sufficient to classify the CNV as disease-causing or pathogenic [58].

However, when studying this region in the ClinGen Dosage Sensitivity Map, we did not find any gene or region curated for triplosensitivity. This means that there is no consensus among ClinGen specialists regarding whether any elements of this region present triplosensitivity. This finding highlights the need for continuous research to reveal novel disease-associated regions of the human genome. In these cases, other criteria must be considered to analyze and classify CNVs, especially when there is no consensus about their impact.

ClinGen, in conjunction with the American College of Medical Genetics and Genomics, has developed a point-based scoring metric for CNV classification. This approach assigns points for different parameters observed, including the number of genes involved, the genomic content, dosage sensitivity predictions and curations, literature support, isolated case reports with the same alteration, clinical phenotype presented by the patient, and inheritance pattern [58]. An online, publicly available CNV classification calculator based on these scoring metrics is available (cnvcalc.clinicalgenome.org/cnvcalc) to facilitate the analysis (Fig. 9.1c). This structural approach helps to classify CNVs even if the chromosomal region they involve is not clearly known to be associated with a genetic disease.

Using the proposed scoring system and considering the number of genes involved, absence of CNV in controls, presence of affected individuals in the DECIPHER database and reports in the literature, the triplication of chromosome 3 presented by this patient was classified as pathogenic, and her final diagnosis was established.

9.6 Closing Remarks

It is currently well established that CNVs are a major source of genetic diversity in humans. Despite numerous technological advances and initiatives to produce CNV maps of the human genome, there are still many unanswered questions on the impact of CNVs and their involvement in human disease. Clinical investigation of CNVs is an essential tool for the accurate diagnosis of several diseases and has a significant impact on the management of patients with complex diseases.

References

1. Alkan C, Coe B, Eichler E. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76.
2. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014a Jan;42(Database issue):D986–92. https://doi.org/10.1093/nar/gkt958.
3. Buchanan J, Scherer S. Contemplating effects of genomic structural variation. Genet Med. 2008;10:639–47.
4. Hastings P, Lupski J, Rosenberg S, et al. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10:551–64.
5. Chiruvella KK, Liang Z, Wilson TE. Repair of double-strand breaks by end joining. Cold Spring Harb Perspect Biol. 2013 May 1;5(5):a012757.
6. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015 Sep;14(5):305–14. https://doi.org/10.1093/bfgp/elv014.
7. Lauer S, Gresham D. An evolving view of copy number variants. Curr Genet. 2019 Dec;65(6):1287–95.
8. Sturtevant AH. The effects of unequal crossing over at the Bar locus in drosophila. Genetics. 1925;10(2):117–47.
9. Bridges CB. The Bar “gene” a duplication. Science. 1936;83(2148):210–1.
10. Muller HJ. Bar duplication. Science. 1936;83(2161):528–30.
11. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, Jorde LB, Posukh OL, Sahakyan H, Watkins WS, Yepiskoposyan L, Abdullah MS, Bravi CM, Capelli C, Hervig T, Wee JT, Tyler-Smith C, van Driem G, Romero IG, Jha AR, Karachanak-Yankova S, Toncheva D, Comas D, Henn B, Kivisild T, Ruiz-Linares A, Sajantila A, Metspalu E, Parik J, Villems R, Starikovskaya EB, Ayodo G, Beall CM, Di Rienzo A, Hammer MF, Khusainova R, Khusnutdinova E, Klitz W, Winkler C, Labuda D, Metspalu M, Tishkoff SA, Dryomov S, Sukernik R, Patterson N, Reich D, Eichler EE. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015a Sep 11;349(6253):aab3761.
12. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, MJP C, Walter K, Meiers S, Kashin S, Garrison E, Auton A, HYK L, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, SA MC, 1000 Genomes Project Consortium, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. Nature. 2015b Oct 1;526(7571):75–81.
13. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. Am J Hum Genet. 2009 Apr;84(4):524–33.
14. Lejeune J, Gautier M, Turpin R. Etude des chromosomes somatiques de neuf enfants mongoliens [study of somatic chromosomes from 9 mongoloid children]. C R Hebd Seances Acad Sci. 1959 Mar 16;248(11):1721–2.
15. Hickey F, Hickey E, Summar KL. Medical update for children with down syndrome for the pediatrician and family practitioner. Adv Pediatr. 2012;59(1):137–57.
16. Huang H, Chen J. Chromosome bandings. Methods Mol Biol. 2017;1541:59–66.
17. Cui C, Shu W, Li P. Fluorescence in situ hybridization: cell-based genetic diagnostic and research applications. Front Cell Dev Biol. 2016 Sep 5;4:89.

18. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 2002 Jun 15;30(12):e57.
19. Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350–4.
20. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 1986;51:263–73.
21. Baldan F, Passon N, Burra S, Demori E, Russo PD, Damante G. Quantitative PCR evaluation of deletions/duplications identified by array CGH. Mol Cell Probes. 2019 Aug;46:101421.
22. Mazaika E, Digital HJ, Droplet PCR. CNV analysis and other applications. Curr Protoc Hum Genet. 2014 Jul 14;82:7.24.1–13.
23. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008 Oct;40(10):1166–74.
24. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009 Nov;6(11 Suppl):S13–20.
25. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007 Oct 19;318(5849):420–6. https://doi.org/10.1126/science.1149504.
26. Shaikh TH. Copy number variation disorders. Current Genetic Medicine Reports. 2017;5:183–90.
27. Online Mendelian Inheritance in Man, OMIM® (n.d.). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD).
28. Bird TD. Charcot-Marie-tooth (CMT) hereditary neuropathy overview. Seattle (WA): University of Washington, Seattle; 1998. Sep 28 [Updated 2020 May 14]. In: Adam MP, Ardinger HH, Pagon RA, et al. editors. GeneReviews® [Internet]. p. 1993–2020.
29. Martinotti A, Cariani CT, Melani C, Sozzi G, Spurr NK, Pierotti MA, Colombo MP. Isolation and mapping to 17p12-13 of the human homologous of the murine growth arrest specific Gas-3 gene. Hum Molec Genet. 1992;1:331–4.
30. De Jong JGY. Over families met hereditarie disposite tot het optreten van neuritiden, gecorreleard met migraine. Monatsschr Psychiatr Neurol. 1947;50:60–76.
31. van Paassen BW, van der Kooi AJ, van Spaendonck-Zwarts KY, Verhamme C, Baas F, de Visser M. PMP22 related neuropathies: Charcot-Marie-tooth disease type 1A and hereditary neuropathy with liability to pressure palsies. Orphanet J Rare Dis. 2014a;9:38. Curr Genet Med Rep. 2017 Dec; 5(4): 183–190
32. Henley SJ, Ward EM, Scott S, et al. Annual report to the nation on the status of cancer, part I: national cancer statistics. Cancer. 2020;126(10):2225–49.
33. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. [published correction appears in J Natl Cancer Inst. 2011 Apr 20;103(8):699]. J Natl Cancer Inst. 2011;103(2):117–28.
34. Petrucelli N, Daly MB, Pal T. BRCA1- and BRCA2-associated hereditary breast and ovarian cancer. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1998. [Updated 2016 Dec 15]. p. 1993–2020.
35. Venkitaraman AR. Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell. 2002;108:171–82.
36. Judkins T, Rosenthal E, Arnell C, et al. Clinical significance of large rearrangements in BRCA1 and BRCA2. Cancer. 2012;118(21):5210–6.
37. Brown MA, Lo LJ, Catteau A, et al. Germline BRCA1 promoter deletions in UK and Australian familial breast cancer patients: identification of a novel deletion consistent with BRCA1:psiBRCA1 recombination. Hum Mutat. 2002;19(4):435–42.

38. Smith LD, Tesoriero AA, Ramus SJ, et al. BRCA1 promoter deletions in young women with breast cancer and a strong family history: a population-based study. Eur J Cancer. 2007;43(5):823–7. https://doi.org/10.1016/j.ejca.2007.01.011.
39. Mai CT, Isenburg JL, Canfield MA, et al. National population-based estimates for major birth defects, 2010-2014. Birth Defects Res. 2019;111(18):1420–35. https://doi.org/10.1002/bdr2.1589.
40. The United Nations. Convention on the rights of persons with disabilities. Treaty Series. 2006;2515:3.
41. Williams JC, Barratt-Boyes BG, Lowe JB. Supravalvular aortic stenosis. Circulation. 1961;24:1311–8.
42. Ewart AK, Morris CA, Atkinson D, Jin W, Sternes K, Spallone P, Stock AD, Leppert M, Keating MT. Hemizygosity at the elastin locus in a developmental disorder. Williams syndrome Nature Genet. 1993;5:11–6.
43. Burn J. Williams syndrome. J Med Genet. 1986;23:389–95.
44. Morris CA, Syndrome W. Updated 2017 mar 23. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1999 Apr 9. p. 1993–2020.
45. Morris CA, Mervis CB, Hobart HH, Gregg RG, Bertrand J, Ensing GJ, Sommer A, Moore CA, Hopkin RJ, Spallone PA, Keating MT, Osborne L, Kimberley KW, Stock AD. GTF2I hemizygosity implicated in mental retardation in Williams syndrome: genotype-phenotype analysis of five families with deletions in the Williams syndrome region. Am J Med Genet. 2003;123A:45–59.
46. Stock AD, Spallone PA, Dennis TR, Netski D, Morris CA, Mervis CB, Hobart HH. Heat shock protein 27 gene: chromosomal and molecular location and relationship to Williams syndrome. Am J Med Genet A. 2003;120A:320–5.
47. Collins RT 2nd, Kaplan P, Somes GW, Rome JJ. Long-term outcomes of patients with cardiovascular abnormalities and Williams syndrome. Am J Cardiol. 2010;105:874–8.
48. Tassabehji M, Hammond P, Karmiloff-Smith A, Thompson P, Thorgeirsson SS, Durkin ME, Popescu NC, Hutton T, Metcalfe K, Rucka A, Stewart H, Read AP, Maconochie M, Donnai D. GTF2IRD1 in craniofacial development of humans and mice. Science. 2005;310:1184–7.
49. Lejeune J, Lafourcade J, Berger R, Vialatta J, Boeswillwald M, Seringe P, Turpin R. Trois ca de deletion partielle du bras court d'un chromosome 5. C R Hebd Seances Acad Sci. 1963;257:3098.
50. Medina M, Marinescu RC, Overhauser J, Kosik KS. Hemizygosity of delta-catenin (CTNND2) is associated with severe mental retardation in cri-du-chat syndrome. Genomics. 2000;63:157–64.
51. Wu Q, Niebuhr E, Yang H, Hansen L. Determination of the 'critical region' for cat-like cry of cri-du-chat syndrome and analysis of candidate genes by quantitative PCR. Eur J Hum Genet. 2005;13(4):475–85.
52. Mefford HC, Batshaw ML, Hoffman EP. Genomics, intellectual disability, and autism. N Engl J Med. 2012 Feb 23;366(8):733–43. https://doi.org/10.1056/NEJMra1114194.
53. Muhle RA, Reed HE, Stratigos KA, Veenstra-VanderWeele J. The emerging clinical neuroscience of autism Spectrum disorder: a review. JAMA Psychiat. 2018;75(5):514–23. https://doi.org/10.1001/jamapsychiatry.2017.4685.
54. Xu J, Zwaigenbaum L, Szatmari P, Scherer SW. Molecular cytogenetics of autism. Curr Genomics. 2004;5:347–64. Front Cell Neurosci. 2019; 13: 57
55. Rylaarsdam L, Guemez-Gamboa A. Genetic causes and modifiers of autism Spectrum disorder. Front Cell Neurosci. 2019 Aug 20;13:385.
56. Schaefer G, Mendelsohn N. Clinical genetics evaluation in identifying the etiology of autism spectrum disorders: 2013 guideline revisions. Genet Med. 2013;15:399–407.
57. Celestino-Soper PB, Violante S, Crawford EL, et al. A common X-linked inborn error of carnitine biosynthesis may be a risk factor for nondysmorphic autism. Proc Natl Acad Sci U S A. 2012;109(21):7974–81.
58. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, Raca G, Ritter DI, South ST, Thorland EC, Pineda-Alvarez D, Aradhya S, Martin CL. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus rec-

ommendation of the American College of Medical Genetics and Genomics (ACMG) and the clinical genome resource (ClinGen). Genet Med. 2020 Feb;22(2):245–57.

59. Boada R, Janusz J, Hutaff-Lee C, Tartaglia N. The cognitive phenotype in Klinefelter syndrome: a review of the literature including genetic and hormonal factors. Dev Disabil Res Rev. 2009;15(4):284–94.
60. Shapira SK, McCaskill C, Northrup H, Spikes AS, Elder FFB, Sutton VR, Korenberg JR, Greenberg F, Shaffer LG. Chromosome 1p36 deletions: the clinical phenotype and molecular characterization of a common newly delineated syndrome. Am J Hum Genet. 1997;61:642–50.
61. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172–83. https://doi.org/10.1038/nrg3871. Epub 2015 Feb 3. PMID: 25645873.
62. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. Nature. 2006;444(7118):444–54. https://doi.org/10.1038/nature05329. PMID: 17122850; PMCID: PMC2669898.
63. Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res. 2007;17(9):1266–77. https://doi.org/10.1101/gr.6557307. Epub 2007 Jul 31. PMID: 17666543; PMCID: PMC1950895.
64. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949–51. https://doi.org/10.1038/ng1416. Epub 2004 Aug 1. PMID: 15286789.
65. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525–8. https://doi.org/10.1126/science.1098918. PMID: 15273396.
66. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85–97. https://doi.org/10.1038/nrg1767. PMID: 16418744.
67. Andy Itsara, Gregory M. Cooper, Carl Baker, Santhosh Girirajan, Jun Li, Devin Absher, Ronald M. Krauss, Richard M. Myers, Paul M. Ridker, Daniel I. Chasman, Heather Mefford, Phyllis Ying, Deborah A. Nickerson, Evan E. Eichler. Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. The American Journal of Human Genetics 2009;84(2):148–61
68. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. De novo rates and selection of large copy number variation. Genome Res. 2010;20(11):1469–81.
69. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J; Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704–12. https://doi.org/10.1038/nature08516. Epub 2009 Oct 7. PMID: 19812545; PMCID: PMC3330748.
70. Lupski JR. Structural variation in the human genome. N Engl J Med. 2007;356(11):1169–71. https://doi.org/10.1056/NEJMcibr067658. PMID: 17360997.
71. Taylor JS, Raes J. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet. 2004;38:615–43. https://doi.org/10.1146/annurev.genet.38.072902.092831. PMID: 15568988.
72. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, Bilous M, Ellis IO, Fitzgibbons P, Hanna W, Jenkins RB, Press MF, Spears PA, Vance GH, Viale G, McShane LM, Dowsett M. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. Arch Pathol Lab Med. 2018;142(11):1364–82. https://doi.org/10.5858/arpa.2018-0902-SA. Epub 2018 May 30. PMID: 29846104.

Chapter 10
The Human Mitochondrial DNA

Regina Célia Mingroni-Netto

10.1 The Origin and the Structure of the Human Mitochondrial DNA (mtDNA)

Mitochondria are cell organelles present in almost all eukaryotic cells and are covered by a double layer of membranes. Structurally, they have four compartments: the outer membrane, the inner membrane, the intermembrane space and the matrix, the region inside the inner membrane. They are the only structures of the animal cells, besides the nucleus, that contain DNA, the mtDNA. In addition, they have their own machinery for the synthesis of RNA and proteins. In plant cells and algae cells, chloroplasts are organelles that also have their own DNA molecules [1]. Mitochondria are dynamic structures, since they are frequently observed as changing their size and shape, or undergoing processes of fusion or fission. The number of mitochondria is also variable among different cell types and tissues, and this number can vary as a response to certain stimuli, for example, frequent muscle contractions [2, 3].

Mitochondria perform many tasks such as pyruvate oxidation, the Krebs cycle and the metabolism of amino acids. Mitochondria also harbor the fatty acid (FA) oxidation machinery, producing acetyl-coenzyme A (acetyl-CoA). They are also key components in calcium signaling, steroid synthesis and apoptosis (programmed cell death), but their outstanding role is certainly the generation of energy as adenosine triphosphate (ATP), by means of the oxidative phosphorylation system (OXPHOS), which occurs via the electron transport chain (ETC).This process requires five protein complexes, four of which make up the mitochondrial

R. C. Mingroni-Netto (✉)
Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, São Paulo, Brazil
e-mail: renetto@ib.usp.br

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_10

respiratory chain (complexes I, II, III and IV) and they are involved in the transport of electrons through the complexes until their final acceptor, molecular oxygen. The four complexes are organized within the inner mitochondrial membrane. The transfer of electrons generates a proton gradient across the inner mitochondrial membrane, where the complex V is also embedded. Complex V is also known as ATP synthase, since it synthesizes ATP via chemiosmotic coupling with the ETC [reviews in [2, 4].

One of the most outstanding ideas about eukaryotic cell evolution was the proposal that mitochondria originated from endosymbiotic bacteria that were incorporated into the eukaryotic cells. Bacteria colonized primordial eukaryotic cells that lacked the ability to use oxygen in energy production. A symbiotic relationship must have become permanent and engulfed bacteria evolved into the present mitochondria. Although the idea was firstly proposed near 1890, the hypothesis was reintroduced and strongly diffused by Lynn Margulis in 1967 [5]. This proposal is referred in the literature as the "endosymbiotic theory for the origin of mitochondria" or "endosymbiotic hypothesis". In accordance, all present genomic evidence points to all mitochondrial genomes, present in all living organisms, as originating from one single endosymbiotic event, probably involving an aerobic alpha-proteobacterium, and they share only one common ancestor, a circular bacterial genome [6–8]. This cell fusion event was estimated to have occurred around 1.5–2 billion years ago. Many aspects of the structure and functioning of mtDNA reinforce the theory of its bacterial origin, such as its circular organization, its presence in multiple copies within the cell, and gene expression mediated by the transcription of large polycistronic RNAs. The same hypothesis of origin from endosymbiotic organisms was applied to explain the origin of chloroplasts in photosynthesizing organisms, organelles that also have their own DNA molecules.

Presently, mtDNA molecules in different species have different sizes and coding capacities, since all of them lost substantial amounts of genetic information, when compared to the coding capacity of the genomes of presently existing bacteria. Most of the genetic information within the ancestral mtDNA was transferred to nuclear chromosomes at different rates and amounts in different species, reducing the genetic independence of mitochondria [9].

In addition to such ancient transfer of sequences from mtDNA to nuclear DNA, there has also been documented the evolutionarily recent transfer of mitochondrial sequences to the nuclear genome. Analysis of the human reference genome sequence shows hundreds of nuclear sequences that are imperfect copies of mtDNA sequences, with varied sizes and locations. These transferred mtDNA sequences usually show inactivating mutations, which pose restrictions to their genetic expression, and are described as nuclear mtDNA sequences, or NUMTs. Some NUMT sequences are present in some individuals, but not in others, thus constituting insertion/deletion polymorphisms in human populations [9].

The first human complete "genome" sequenced was that of the mitochondria in 1981 [10], by Fred Sanger and colleagues at Cambridge, many years before the Human Genome Project began. It was subsequently referred as the Cambridge

Reference Sequence (CRS). The nucleotide numbering of mtDNA sequence presently in use is based on a revised and corrected version of this reference, the rCRS [10, 11]. The human mtDNA comprises 16,569 bp. It is a double-stranded DNA molecule, that resembles bacterial genomes because it is densely packed with genes, and circular. One of the strands is called the heavy strand (HS) because it is guanine-rich compared to the light strand (LS), which is cytosine-rich.

MtDNA has a small non-coding region, the 1.1 kb displacement loop (D loop), also named as control region, which includes elements that regulate transcription and replication: the two major transcription initiation sites that are required to generate polycistronic transcripts (HSP1 and HSP2) and one of the origins of mtDNA replication, the one of the heavy strand (OH). Parts of the control region are variable in sequence and are referred as hypervariable segments (HVS) I, II and III. The second origin of replication, on the light strand (OL), is outside the control region and is located near 11 kb away from OH. The structure of human mtDNA is schematically represented in Fig. 10.1.

The circular molecular also comprises a larger coding region containing 37 genes. Thirteen genes encode 13 different proteins synthesized by mitochondrial ribosomes, and related to the ETC. All 13 proteins act as subunits of the mitochondrial enzyme complexes involved in oxidative phosphorylation (OXPHOS). Twenty-two genes are templates for the transcription of 22 transfer RNAs (tRNA) which act exclusively in the translation of mitochondrial peptides. Finally, two genes are for two ribosomal RNA (rRNA) molecules, 12S and 16S, components of mitochondrial ribosomes.

The mitochondrial ribosomes have a sedimentation coefficient of 55S and are constituted by two subunits, 39S and 28S, in which are present the ribosomal RNA molecules of 16S (expressed by *MT-RNR2*) and 12S (expressed by *MT-RNR1*), respectively. There are no introns in the human mtDNA and more than 90% of the mitochondrial 'genome' specifies a protein or a functional RNA. The sequences of neighboring genes are continuous or separated by only a few non-coding bases, and there is extensive overlap of coding sequences between the two strands of the circle, the heavy (H) strand and the light strand (L).

It is important to highlight that there are estimates that, in human mitochondria, a proteome of 1100–1700 different proteins with varied functions is acting. Thus, the capacity of mtDNA of coding only 13 mitochondrial peptides reveals that, presently, mitochondrial functions are largely dependent on proteins encoded by nuclear genes, translated in cytoplasmic ribosomes, and imported by mitochondria. In accordance, only 13 of the 80 proteins required for oxydative phosphorylation are coded by the mitochondrial genome. Crucial proteins needed for replication, transcription and repair of mtDNA are encoded by nuclear genes. As a consequence, mitochondria are under dual genetic control, by its own DNA and the nuclear genome.

The mtDNA is comparatively protein-free, as are bacterial genomes, because it is not condensed with histones as nuclear chromosomes are. Nevertheless, it is packed with some proteins to form nucleoids, nucleoprotein structures that are associated with the inner mitochondrial membrane. The mitochondria nucleoids contain the protein machinery required for DNA replication, transcription, repair and

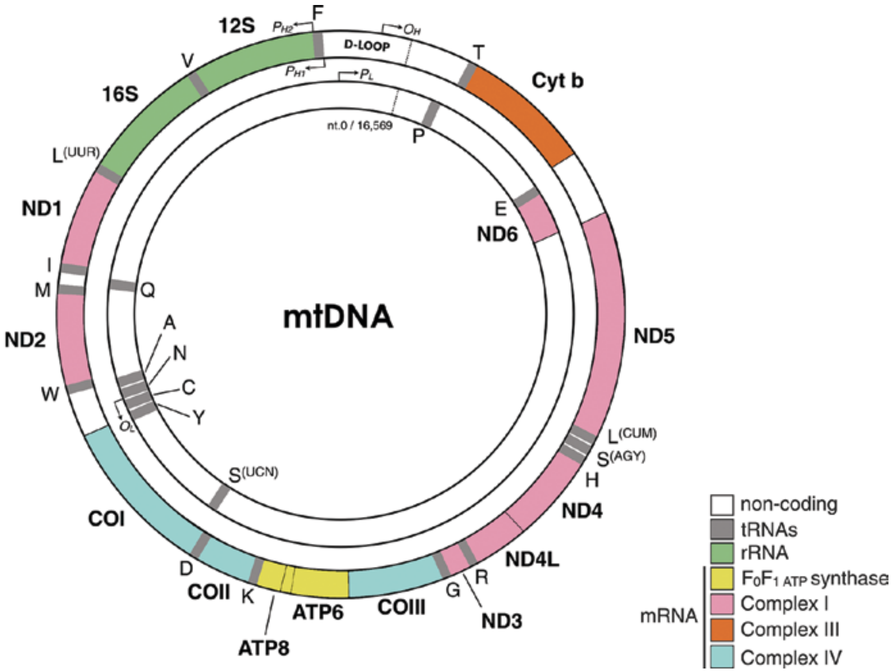


Fig. 10.1 The mtDNA, containing 37 genes (Figure adapted from Picard et al. [3]). O_H = Origin of replication of the heavy strain; O_L = Origin of replication of light chain; P_L = promoter of the light strain; P_{H1} = Promoter 1 of heavy strain; P_{H2} = Promoter 2 of the heavy strain; List of genes and their products: *Cyt b* Mitochondrially encoded cytochrome b (*MT-CYB*), *ND6* Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 6 (*MT-ND6*), *ND5* Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 5 (*MT-ND5*), *ND4* Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 4 (*MT-ND4*), *ND4L* Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 4 L (*MT-ND4L*), *ND3* Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 3 (*MT-ND3*), *ND2* Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 2 (*MT-ND2*), *ND1* Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 1 (*MT-ND1*), *COIII* Mitochondrially encoded cytochrome c oxidase III (*MT-CO3*), *ATP6* Mitochondrially encoded ATP synthase membrane subunit 6 (*MT-ATP6*), *ATP-8* Mitochondrially encoded ATP synthase membrane subunit 8 (*MT-ATP8*), *COII* Mitochondrially encoded cytochrome c oxidase II (*MT-CO2*), *COI* Mitochondrially encoded cytochrome c oxidase I (*MT-CO1*), *16S* Mitochondrially encoded 16S rRNA (*MT-RNR2*), *12S* Mitochondrially encoded 12S rRNA (*MT-RNR1*), *P* Mitochondrially encoded tRNA-Pro (CCN) (*MT-TP*), *T* Mitochondrially encoded tRNA-Thr (ACN) (*MT-TT*), *E* Mitochondrially encoded tRNA-Glu (GAA/G) (*MT-TE*), *L* (CUM): Mitochondrially encoded tRNA-Leu (CUN) 2 (*MT-TL2*), *S* (AGY): Mitochondrially encoded tRNA-Ser (AGU/C) 2 (*MT-TS2*), *H* Mitochondrially encoded tRNA-His (CAU/C) (*MT-TH*), *R* Mitochondrially encoded tRNA-Arg (CGN) (*MT-TR*), *G* Mitochondrially encoded tRNA-Gly (GGN) (*MT-TG*), *K* Mitochondrially encoded tRNA-Lys (AAA/G) (*MT-TK*), *D* Mitochondrially encoded tRNA-Asp (GAU/C) (*MT-TD*), *S* (UCN): Mitochondrially encoded tRNA-Ser (UCN)1 (*MT-TS1*), *Y* Mitochondrially encoded tRNA-Tyr (UAU/C) (*MT-TSY*), *C* Mitochondrially encoded tRNA-Cys (UGU/C) (*MT-TC*), *N* Mitochondrially encoded tRNA-Asn (AAU/C) (*MT-TN*), *A* Mitochondrially encoded tRNA-Ala (GCN) (*MT-TA*), *W* Mitochondrially encoded tRNA-Trp (UGA/G) (*MT-TW*), *M* Mitochondrially encoded tRNA-Met (AUA/G) (*MT-TM*), *Q* Mitochondrially encoded tRNA-Gln (CAA/G) (*MT-TQ*), *I* Mitochondrially encoded tRNA-Ile (AUU/C) (*MT-TI*), *L (UUR)* Mitochondrially encoded tRNA-Leu (UUA/G) 1 (*MT-TL1*), *V* Mitochondrially encoded tRNA-Val (GUN) (*MT-TV*), *F* Mitochondrially encoded tRNA-Phe (UUU/C) (*MT-TF*)

packaging of mtDNA, including the mtDNA polymerase, POLG (or POL gamma) and the main mtDNA transcription factor (TFAM; mitochondrial transcription factor A), as well as mtDNA helicases (TWINKLE) and other proteins, such as single strand DNA binding proteins. The major structural component of mitochondrial nucleoids is TFAM, an abundant protein involved in mtDNA transcription, acting as a mitochondrial transcription factor besides its function of packing the mtDNA into nucleoids. Some microscopy experiments suggested that each nucleoid is 100 nm in diameter and comprises one copy of mtDNA packaged with multiple TFAM molecules, but there are references to one nucleoid containing more than one copy of the mtDNA molecule [2].

The replication of both strands of the human mtDNA, heavy (H) and light (L) chains, is unidirectional and starts at specific origins, OH and OL, respectively. Although mtDNA is generally double-stranded, the repeated replication of a small segment of the H strand results in the production of a shorter third strand called 7SDNA. The 7S DNA can pair with the L-strand, displacing the H strand, which then forms a small loop, called displacement loop or D-loop. As shown in Fig. 10.1, this region also contains the major promoter regions and the origin of replication for the H-strand (OH), and this explains why it is referred as the CR/D-loop region, or the control Region. The three main factors in mtDNA replication are DNA polymerase gamma (POLG), the mitochondrial helicase (TWINKLE) and the mitochondrial single-strand DNA binding protein (mtSSB). The POLG holoenzyme is a heterotrimer that consists of two identical subunits (PolG-B) and one catalytic subunit named PolG-A. The DNA helicase forms a hexameric structure and is required at the replication fork where it unwinds the double-stranded DNA ahead of POLG to expose the template for replication. POLG was thought to be the sole DNA polymerase in mitochondria, being involved in replication and repair, but at least four other mitochondrial polymerases were found to be related to mtDNA maintenance and repair.

The exact mechanism of mtDNA replication is unknown and there are conflicting theories to explain the process. According to one of the models of mtDNA replication, only after about 2/3 of the H strand is replicated from the OH, the replication machinery reaches the origin of replication of the L strand (OL), starting the replication of this strand in the opposite direction. This model of replication was the first to be proposed, but evidence for an alternate model of replication was also obtained, in which leading-lagging strand DNA replication is coordinated, suggesting simultaneous replication of both strands, similar to the replication strategy observed in linear nuclear chromosomes. The controversy about the two possible modes of DNA replication remains unsolved [2, 4].

MtDNA replication does not seem to be subject to a strict control of copy number. The replication rates are flexible and seem to vary. Between 100 and 10,000 mtDNA copies may be found in the inner mitochondrial compartment (matrix) in different cell types, and oocytes are estimated to have 100,000 mtDNA copies. The mitochondrial genome is frequently renewed and replication does not occur in a specific phase of cell cycle, as it happens with nuclear DNA. The number of copies of mtDNA is observed to vary with cell type and within the same cell type, and it can change in accordance to energy demands.

Mitochondria are also equipped with mechanisms to repair damaged DNA, but not all components of nuclear DNA repair have their counterparts in mitochondria. Base excision repair (BER) is active within mitochondria and repairs small lesions such as alkylated and oxidized bases. There are seven different glycosylases that initiate BER in mtDNA and there is evidence that DNA synthesis after damage removal is performed by polymerase beta. Proteins involved in recombination and translesion DNA synthesis were also identified within mitochondria, as well as some proteins related to nucleotide excision repair (NER), although the existence of this repair pathway remains uncertain.

The expression of mtDNA is widely different from nuclear genes. The two strands of mtDNA are transcribed to give two long polycistronic transcripts that resemble bacterial polycistronic RNAs. The long transcripts are cleaved to generate individual mRNAs or functional RNAs such as tRNAs and rRNAs. Transcription initiates from three possible promoters: one located in the light-strand (LSP) and two located on the heavy strand (HSP1 and HSP2). The HSP1 enables the transcription of the 12S and 16S rRNA genes while the HSP2 enables the transcription of the entire H-strand as a polycistronic transcript. The L-strand is transcribed from the light strand promoters (LSP). Transcription is, thus, bidirectional and requires TFAM, the mitochondrial transcription factor B2 (TFB2M) and mitochondrial RNA polymerase. The transcription factors assemble at the promoters to initiate the synthesis of the polycistronic RNAs that are later processed into smaller RNAs. TFAM appears to be regulated via post-translational modifications and these modifications may be related to epigenetic mechanisms of regulation of mitochondrial transcription. It seems that RNA transcription in mitochondria is also regulated by control of the number of copies of mtDNA. An increase in energy need is usually followed by an increase of copy number of mtDNA in tissues, resulting in enhanced expression of mitochondrial genes [2].

The mitochondrial ribosomes exclusively translate peptides coded by mtDNA genes and do not translate nuclear derived mRNAs. As a result, the mitochondrial genetic code could drift in evolution from the “Universal Genetic Code”. The mitochondrial code is slightly different from the genetic code that is used in cytoplasmic ribosomes of almost all living organisms. According to the “Universal Genetic Code”, 61 codons specify amino acids and three are stop codons: UAA, UAG and UGA. In the human mitochondrial code, 60 codons specify amino acids and there are four stop codons: UAA, UAG (also stop codons in the nuclear code) and AGA and AGG (which would specify arginine in the nuclear code). UGA, which is a stop codon in the universal code, encodes tryptophan in mitochondria. AUA encodes methionine in mitochondria, instead of isoleucine, as in the conventional code.

The mitochondrial genome shows unique genetic characteristics, such as matrilineal inheritance, lack of recombination, and high sequence variability, which make it distinct from the nuclear genome in many genetic features detailed in the next topics of this review.

10.2 The Mutation Rates in mtDNA are Elevated When Compared to Nuclear DNA

A 10–1000 fold higher mutation rate was estimated in mtDNA when compared to the substitutional mutation rate in nuclear DNA, depending on the portion of the mtDNA evaluated. The mutational rate in the control region is the highest, leading to a large number of different mtDNA sequences in human populations. The estimated mutation rates are not uniform, ranging from 2×10^{-7} in some tRNA sequences to 5×10^{-6} in the HVS-I and HVS-II [12].

Several facts account for the observation of elevated rates of mutations. First, mtDNA is constantly attacked by reactive oxygen species (ROS) generated by oxidative phosphorylation, because of its close proximity to the components of the respiratory chain. ROS are potent genotoxic agents and they are responsible for higher nucleotide instability. In spite of being packed in mitochondrial nucleoids, with some DNA-associated proteins such as TFAM and polymerases, mtDNA is not as tightly packed with proteins as nuclear chromosomes. Nuclear DNA exists in a complex chromatin structure mainly organized by histones (see Chap. 2), and histones are not present in mitochondria. Thus, it is assumed that mtDNA is less protected than nuclear DNA from genotoxic agents. Some DNA repair pathways that can partly cope with oxidative damage are present in mitochondria, but they are not as complex as those acting in nuclear DNA. Besides, given that the number of copies of mtDNA in each cell is usually much higher than sequences in the nucleus (in which a specific sequence is replicated only once every cell cycle), the replication history of any mtDNA molecule will be longer than a nuclear sequence, since there are more replication events per unit of time. This increases the probability of replication errors. It has also been pointed that mtDNA, because of its unusual mode of replication, spends more of its time in single-stranded form and more exposed to damage, and this is especially true for the D-loop region [2, 12].

10.3 Rare mtDNA Variants Lead to Hereditary Diseases with Maternal Transmission

Mitochondrial diseases is a term used to describe a clinically heterogeneous group of genetic disorders characterized by defective oxidative phosphorylation. They result from dysfunction of mitochondria and may lead to a variety of symptoms that can be detected in neonatal phase, childhood or adulthood. The dysfunction of mitochondria results in a chronic loss of cellular energy or incapacity to meet cellular energy demands. As a consequence, the resulting symptoms can be present in isolated organs, but they often cause multiple system impairment. They are either a consequence of pathogenic variants in nuclear genes encoding protein products that are relevant to mitochondria, or result from pathogenic alterations in mitochondrial genes that code mitochondrial proteins or RNA molecules. Given an estimated proteome of 1100–1700 different proteins, in fact, most proteins involved in mitochondrial metabolism are nuclearly encoded. Nevertheless, a number of human genetic diseases is due to pathogenic alterations in mtDNA and these are exclusively maternally inherited [14]. A woman carrying a mutated mtDNA sequence passes it to all children, but men will not transmit mtDNA to their progeny. We will only review in this chapter diseases that result from pathogenic variants in the mtDNA.

Mutations in mtDNA may affect specific proteins of the respiratory chain, when they occur in protein coding genes, or they may affect the synthesis of mitochondrial proteins as a whole, if they occur in tRNA or rRNA genes. These variants in mtDNA may result in many clinical features and syndromes with overlapping clinical symptoms, variable expressivity and penetrance, thus representing a real challenge for their clinical recognition and accurate classification. Mitochondria are ubiquitous and every tissue can be affected by a mtDNA pathogenic variant, and this is why mitochondrial diseases are usually multisystemic. However, there are some clinical features that are recurrent in many of the mitochondria-related syndromes, since the basic biological defect underlying many of these diseases is impairment of energy production. Thus, mtDNA alterations usually impair more severely organs or tissues with high energy needs, such as the brain, skeletal muscles and heart, and those that have to respond quickly to abrupt changes in the environment, at the expenses of consuming more ATP [4]. Impairment of neurologic functions and of sensorial systems as hearing and vision, muscle weakness, diabetes and other endocrine diseases are frequent features of mitochondrially inherited diseases.

The first report of inherited disease resulting from mutation in the mtDNA was that of Wallace et al. in [13], who reported that mtDNA point mutations and deletions caused MERFF (Myoclonus, Epilepsy and Ragged-red Fibers mitochondrial encephalomyopathy) and this finding was considered a breakthrough for molecular medicine. Since then, it has been established that many inherited and acquired mtDNA defects are at the roots of many pediatric and adult diseases. Many different disorders were described as resulting from mtDNA alterations, which range from single nucleotide substitutions to large mtDNA deletions. Genetic mitochondrial disorders have initially remained in the domain of neurology, but the discovery of a broader range of diseases, with clinically complex phenotypes, have placed mitochondrial diseases in many different medical specialties, such as cardiology, endocrinology, immunology, oncology, and others.

Large-scale mtDNA deletions are usually associated to three main phenotypes: chronic progressive, external opthalmoplegia (PEO), Kearns-Sayre syndrome (KSS) and Pearson syndrome. Pearson syndrome shows the most severe clinical presentation: patients present sideroblastic anemia and pancreatic dysfunction early in life, and the condition may be fatal. KSS patients present with ptosis, PEO and pigmentary retinopathy and may have multisystem impairment including myopathy, ataxia and cardiac conduction defects. Deletions were also described in cases of MELAS, with severe neurological symptoms including encephalopathy and stroke-like episodes. Although large deletions often arise sporadically, they result in devastating syndromes [14, 15].

In contrast to many mitochondrial diseases that are multisystemic, it is puzzling that some mitochondrial variants result only in tissue-specific effects. This is the case of m.1555A > G in the 12S rRNA gene (*MT-RN1*), causative to non-syndromic hearing loss. This and some other hearing loss-associated mtDNA variants located mainly in the *MT-TRN1* and *MT-TS1* genes are rarely accompanied by other clinical features. Many variants in the *MT-TRN1* gene are related to hearing loss, which is anticipated or intensified after administration of aminoglycoside antibiotics, because they increase susceptibility to aminoglycoside-induced hearing loss. This example is iconic in the demonstration of the interaction of mitochondrial functions with environment, and how this interaction affects age of onset, severity and progression of the disease phenotypes. It is widely recognized that the most frequent mtDNA variant associated with non-syndromic hearing loss is m.1555A > G and it was also the first to be described, in 1993 [16]. The penetrance of hearing loss is estimated near 40–50%, in pedigrees that show exclusively maternal inheritance. The age at onset and severity of hearing loss are variable within pedigrees, and besides being correlated to aminoglycoside treatment, it is possibly influenced by other factors, such as nuclear modifier genes [17]. It was speculated by some authors that the m.1555A > G substitution makes the mitochondrial ribosome RNA more similar to the bacterial counterpart, increasing the affinity of the human mitochondrial ribosome to aminoglycosides, and mitochondrial translation is consequently compromised.

A summary of the clinical characteristics of the most frequent mitochondrially inherited diseases, with the corresponding mtDNA variants associated, is presented in Table 10.1. There are no straightforward correlations between genotypes and phenotypes regarding mtDNA variants [14, 18]. In other words, no clear correlations are seen between the site of the mutation and the clinical phenotype, even within the same gene, except for some variants. For instance, variants in the tRNA^{Leu} gene *MT-TL1* (UUA/G) may be associated with mitochondrial encephalopathy, lactic acidosis and stroke like episodes (MELAS) syndrome, but they can be causative to other syndromes. On the other hand, mutations in different genes can cause the same syndrome, and MELAS is one of the examples (Table 10.1). Moreover, it is striking that some small size variants (substitutions) relate to different clinical findings. An estimate based on a cohort study in England pointed to a prevalence of diseases caused by mtDNA pathogenic alterations near 9.6 cases per 100,000.

Given the difficulty in clinical classification of mitochondrial inherited diseases and the lack of clear correlation between specific variants and clinics, the molecular diagnosis of mitochondrial diseases has always been troublesome and expensive. Sanger sequencing of many mitochondrial genes was needed until the causative variant was found [4]. Besides, heteroplasmy, as explained in the next section, has always been a challenge to molecular analysis since, in many patients, the mutated mtDNA lineage is present in very low frequencies in circulating blood or it is detected only in the mostly affected tissues, such as skeletal muscle.

Table 10.1 Some genetic diseases caused by mtDNA variants (based in [18, 35, 37, 46])

Disease	Main clinical features	Mutated genes	Most frequent causative variants
Chronic progressive external opthalmoplegia (CPEO)	Ptosis, ophthalmoparesis, proximal myopathy, exercise intolerance, lactic acidosis and ragged-red fibers on muscle biopsy	Several *MT-TL1* *MT-TI* *MT-TI* *MT-TN* *MT-TN* *MT-TK* *MT-TL2* *MT-TL2* *MT-TL2*	mtDNA deletions m.3243A > G; A > T m.4298G > A m.4308G > A m.5690A > G m.5703G > A m.8344A > G m.12276G > A m.12294G > A m.12315G > A m.12316G > A
Kearns-Sayre syndrome (KSS)	Progressive external opthalmoplegia, ptosis, pigmentary retinopathy, cardiac conduction abnormalities, ataxia, diabetes mellitus, sensorineural hearing loss, myopathy, lactic acidosis and ragged-red fibers on muscle biopsy	Several *MT-TL2*	mtDNA deletions m.12315G > A
Leber hereditary optic neuropathy (LHON)	Bilateral visual failure with optic atrophy, dystonia, cardiac pre-excitation syndromes	*MT-ND1* *MT-ND1* *MT-ND1* *MT-ND1* *MT-ND1* *MT-ND4L* *MT-ND4* *MT-ND5* *MT-ND5* *MT-ND5* *MT-ND6* *MT-ND6* *MT-ND6* *MT-ND6*	m.3460A > G m.3635G > A m.3700G > A m.3733G > A m.4171C > A m.10663 T > C m.11778G > A m.13051G > A m.13094 T > C m.13379A > C m.14482C > A;C > G m.14484 T > C m.14495A > G m.14568C > T
Mitochondrial encephalopathy, lactic acidosis, stroke -like episodes (MELAS)	Stroke-like episodes, encephalopathy, migraine, seizures, myopathy, cardiomyopathy, hearing loss, endocrinopathy including diabetes, ataxia, hemiparesis, cortical blindness, lactic acidosis and ragged-red fibers on muscle biopsy	*MT-TF* *MT-TV* *MT-TV* *MT-TL1* *MT-TL1* *MT-TL1* *MT-TL1* *MT-TL1* *MT-ND1* *MT-TQ* *MT-TM* *MT-ND3* *MT-TH* *MT-ND5* *MT-ND5* *MT-ND5MT-ND5*	m.583G > A m.1630G > A m.1644G > A m.3243A > G m.3256C > T m.3258 T > C m.3260A > G m.3271 T > C m.3697G > A m.4332G > A m.4450G > A m.10158 T > C m.12147G > A m.13094 T > C m.13379A > C m.13513G > A m.13514A > G

Table 10.1 (continued)

Disease	Main clinical features	Mutated genes	Most frequent causative variants
Myoclonus, epilepsy and ragged-red fibers (MERRF)	Stimulus-sensitive myoclonus, seizures, ataxia, cardiomyopathy, lactic acidosis and ragged-red fibers on muscle biopsy	*MT-TK*	m.8340G > A
		MT-TK	m.8344A > G
		MT-TK	m.8356 T > C
		MT-TK	m.8363G > A
		MT-TH	m.12147G > A
Neurogenic weakness with ataxia and retinitis pigmentosa (NARP)	Ataxia, peripheral neuropathy, pigmentary retinopathy, weakness.	*MT-ATP6*	m.8993 T > C; m.8993 T > G
MILS maternally inherited Leigh's syndrome	Seizures, ataxia, psychomotor delay, dystonia, muscle weakness, occasional pigmentary retinopathy, optic atrophy and lactic acidosis	*MT-TV*	m.1644G > A
		MT-ND1	m.3697G > A
		MT-ND1	m.3890G > A
		MT-TM	m.4450G > A
		MT-TW	m.5537_5538insT
		MT-TK	m.8363G > A
		MT-ATP6	m.8851 T > C
		MT-ATP6	m.8993 T > C; T > G
		MT-ATP6	m.9176 T > C
		MP-ATP6	m.9185 T > C
		MT-ND3	m.10158 T > C
		MT-ND3	m.10191 T > C
		MT-ND3	m.10197G > A
		MT-ND4	m.11777C > A
		MT-ND5	m.12706 T > C
		MT-ND5	m.13379A > C
		MT-ND5	m.13514A > G
		MT-ND6	m.14459G > A
		MT-ND6	m.14487 T > C
Non-syndromic hearing loss	Sensorineural hearing loss (some related to aminoglycoside Induced hearing loss)	*MT-RNR1*	m.1027 A > G
		MT-RNR1	m.1291 T > C
		MT-RNR1	m.1494C > T
		MT-RNR1	m.1555A > G
		MT-TL1	m.3243A > G
		MT-TS1	m.7445A > C;A > G
		MT-TS1	m.7465A > C
		MT-TS1	m.7497G > A
		MT-TS1	m7505T > C
		MT-TS1	m.7510 T > C
		MT-TS1	m.7511 T > C
		MT-TS1	m.7512 T > C
		MT-TH	m.12201 T > C

The variants highlighted in gray are the most frequently found among patients with the disease

The introduction of next generation sequencing (NGS or massive parallel sequencing) to the study of mtDNA and its rapid transfer to clinical practice speeded up and increased the precision in the molecular diagnosis of mitochondrial diseases (see below). Besides allowing detection of heteroplasmic mtDNA sequences in low frequencies considerably better than previously possible with conventional techniques (e.g. Sanger sequencing), NGS allowed simultaneous sequencing of all mitochondrial genes and many nuclear genes related to mitochondrial diseases in only one experiment, for instance, after their capture and selection for NGS, constituting a panel including mitochondrial genes and nuclear mitochondrial-disease related genes [19]. Nevertheless, challenges still remain in the clinical recognition of mitochondrial diseases.

10.4 Heteroplasmy Is a Generalized Phenomenon in mtDNA Inheritance and Relates to the Clinical Expression of Diseases

Each cell, depending on the type, may contain hundreds of mitochondria, and each mitochondrion harbors some copies of its genome. Thus, thousands of mtDNA molecules may be present in each cell and pathogenic mtDNA variants may be present in some, but not all of these molecules.

Alterations in the mtDNA sequence occur very frequently, and they can exist, at least transiently, as two or more different molecules with distinct nucleotide sequences within a single mitochondrion, cell, tissue or organism. The multicopy nature of mtDNA thus allows the phenomenon of heteroplasmy, a unique aspect of mitochondrial inheritance. Heteroplasmy is defined as the co-existence, in cells, tissues, organs or individuals, of mtDNA molecules with different nucleotide sequences. Heteroplasmy was firstly recognized as frequent, and clinically relevant, in patients from pedigrees in which mitochondrially inherited diseases were segregating (review in [4]).

The existence of multiple mtDNA copies within a cell greatly affects the impact of pathogenic variants of mtDNA. Wild-type copies of mtDNA encode normal copies of mitochondrial proteins or RNA molecules, while mutated mtDNA encode abnormal products. Besides, mitochondria are continuously involved in fusion and fission, which allow exchange of proteins, RNA and other components between mitochondria located within the same cell. Variability in clinical expression and age at onset, or lack of penetrance, are features of many diseases with classical Mendelian transmission. However, these features are enormously enhanced in hereditary diseases that result from mtDNA mutations, partly because of heteroplasmy. The clinical expression of a pathogenic variant in mtDNA correlates with the proportion of wild-type and mutant copies. In many of the diseases, a minimum amount of mutated mtDNA must be present before any kind of cellular dysfunction occurs and clinical signs of disease become apparent, in a threshold effect. The threshold seems to be lower in tissues that are largely dependent on oxidative metabolism, such as brain, heart, muscle, retina, endocrine glands and kidney, thus explaining why these are frequently compromised in mitochondrial diseases (review in [18]).

The random distribution of mitochondria at the time of cell divisions can lead to fluctuations in the proportion of mutated mtDNA that is received by daughter cells. Whenever a pathogenic threshold is surpassed, the cell phenotype can change. This explains some age-related and some tissue-related variability of clinical symptoms in individuals with mtDNA disorders. The investigation of families in which mtDNA pathogenic variants are segregating has shown that onset of clinical manifestation and severity of disease may be correlated to the frequency of heteroplasmy. For instance, different mutation proportions explain the different degrees of severity of neuropathy, ataxia and retinitis pigmentosa in Leigh's syndrome. Heteroplasmy is probably one of the major explanations for the wide variation of phenotypes between maternally related individuals sharing a mtDNA variant that leads to inherited disease, and it has been investigated for decades in the context of mitochondrial diseases.

On theoretical grounds, heteroplasmy should be expected to be a very common event, since each oocyte contains many mitochondria, each one with many mtDNA copies, and mutation rates are elevated in mitochondria. Indeed, all presently existing inherited mtDNA variants must have existed transiently in heteroplasmic states when they first rose after mutation, before their fixation in the germ cells. However, conventional strategies of molecular assessment of heteroplasmy, for instance, Sanger sequencing, had shown severe sensitivity limitations to detect alternative sequences present in low number of copies.

More recently, the NGS technology provided excellent opportunities to reassess the matter of heteroplasmy, allowing more precise and quantitative approaches of investigation of mutated mtDNA in healthy and abnormal tissues, because of its deep coverage. This has largely confirmed that heteroplasmy is a generalized phenomenon, much more frequent than initially suspected [20–22].

The observation that individuals from the same sibship or pedigree may have different proportions of heteroplasmy always puzzled geneticists. Furthermore, large shifts in the frequency of heteroplasmy can be observed in only one transmission, from mother to child. This also contributes to explain why individuals in the same pedigree show different clinical presentations of the mitochondrial disease, with striking differences in severity, and some hypothesis were proposed to explain the findings. It is now widely recognized that heteroplasmy frequency shifts can be explained by a so-called mitochondrial "bottleneck" [2, 4]. In oogenesis, the population of mitochondria that is present in a mammalian oocyte (near 100,000) results from the amplification of a reduced initial number of mitochondria, containing a small number of mtDNA copies. During female embryogenesis, the primordial germ cells (oogonia) develop and early in this process there is a bottleneck of a few hundred mtDNA copies and, by chance or selection, some mtDNA lineages containing variants may be eliminated. Thus, the rapid amplification of mtDNA from a reduced pool of molecules may result in different oocytes bearing widely different proportions of mutated and wild-type mtDNA, leading to abrupt changes of the frequency of heteroplasmy in one generation. It has also been pointed that, in mice, a subsequent bottleneck occurs in the early postnatal period. Fertilized oocytes develop into blastocysts by day 4.5 after conception and this step is known as preimplantation development. The mtDNA copy number remains constant during the preimplantation period while cells divide. It was proposed that, in this period, there is a reduced number of mtDNA molecules per cell and unequal segregation mtDNA molecules in the following cell divisions, widening differences in the frequency of heteroplasmic variants.

Nevertheless, one must always bear in mind thaty is far from being the sole explanation for differences in expressivity of mitochondrial disorders. Some mtDNA variants, for instance, m.1555A > G that is causative of hearing loss, has been detected in in most of the pedigrees in which it was found. Age of onset and severity of hearing loss are extremely variable within these pedigrees and penetrance of hearing loss hardly exceeds 50% [17]. In other diseases, clinical and phenotypic variability may exist among patients with similar levels of heteroplasmy. Many other factors may affect clinical variability and progression of mitochondrial diseases and these are probably environmental factors. As previously mentioned in the case of m.1555A > G, the penetrance of hearing loss is known to be strongly influenced by the administration of aminoglycosides. Lifestyle, exercise, smoking, exposure to oxidant molecules and aging are the most likely environmental modifiers of the progression of mitochondrial disease. Besides, since the majority of mitochondria proteome is coded by nuclear genes, variability in genotypes in nuclear genes that code mitochondrial components must have a role in phenotypic expression of mitochondrial dysfunction. In fact, attempts were made to map nuclear modifiers of expression of mtDNA related diseases. It also remains plausible that single nucleotide polymorphisms (SNPs) in the mitochondrial genome, for instance, polymorphic sites associated to the definition of haplogroups, may have a subtle effect on manifestation of mitochondrial diseases.

Although molecular diagnosis has recently improved its accuracy, genetic counseling of families with mtDNA pathogenic variants remains a complex task, since it is almost impossible to predict precisely the recurrence risk or severity of a disease in the following generation, given the many uncertainties regarding the prediction of heteroplasmy levels and other factors that affect the clinical outcome [18, 19].

10.5 Human mtDNA Is Maternally Transmitted, But Striking Exceptions Have Been Described

The transmission of human mtDNA is strictly maternal. All mtDNA in the zygote derive from the ovum. Therefore, a mother carrying a mtDNA variant in homoplasmy passes it to all her children, but only their daughters will transmit it to progeny. In other words, no one is expected to inherit mtDNA from the father. Although maternal inheritance was long ago assumed as a dogma in Biology, the cellular mechanisms underlying this assumption came to light more recently.

Although some basic facts about the biology of human and mammalian fertilization are known since the decade of 1930, unfortunately, many biology textbooks still replicate a wrong concept to explain the exclusively maternal inheritance of mtDNA: that sperm midpiece and tail do not enter oocytes in mammalian fertilization, thus excluding paternal mitochondria from the zygote. Although this happens in some exceptional cases, such as in the Chinese hamster, this was equivocally extrapolated to other mammals and humans. This is clearly a misconception and it is surprising that it has survived so long, since it has been long demonstrated that sperm fully enters the oocytes in fertilization, and sperm mitochondria are indeed observed within the oocytes after fertilization. In mammalian fertilization, the mitochondria-rich midpiece of the sperm tail enters the oocyte and it is a fact that sperm contributes to the fertilized oocyte's pool of mitochondria. Tail and midpiece can be traced within the zygote for several division cycles after fertilization [23].

The number of mtDNA molecules within a single spermatozoon is certainly much lower than the copy number in oocytes. It was estimated that a human spermatozoon may contain something in between 100 and 1200 mtDNA copies, while a human oocyte contains between 100,000 and 250,000 molecules. This effect of "dilution" of paternal mtDNA allows to predict a reduced contribution of paternal mtDNA in offspring [25]. However, even under such scenario, one occasionally would see transmission of paternal mtDNA, but these exceptions are extremely rare. Thus, additional mechanisms that halt paternal mtDNA transmission were expected to be revealed.

Near the 1960s, it was already known that in rat, all sperm structures penetrate the oocyte, but in the first cell divisions after fertilization, paternal mitochondria swell, lose their cristae and disintegrate, being completely eliminated in the preimplantation embryos. In a remarkable contribution, Sutovsky and co-workers [24] provided the first evidence of the molecular mechanisms acting on the elimination of paternal mitochondria after fertilization in mammals. They demonstrated that ubiquitination of sperm mitochondria during mammalian spermatogenesis is a key factor that leads to elimination of sperm mitochondria by means of proteolysis in the oocyte cytoplasm. Poly-ubiquitination is one of the cellular processes in which a protein is tagged to proteolysis. This is achieved by a covalent binding of the ubiquitin peptide to lysine residues of the targeted proteins. Mammalian mitochondria are ubiquitinated probably in the male reproductive tract and ubiquitin tagging of the sperm mitochondrial membranes culminates in their recognition by the ubiquitin-proteasome-dependent proteolytic machinery. According to Sutovsky and co-workers [25], the mitochondria inner membrane protein prohibitin would be the best candidate as an ubiquitin substrate. This serves as a death sentence for paternal mitochondria after fertilization, since it triggers their elimination. During mammalian spermatogenesis, mitochondrial ubiquitination is already detected at the secondary spermatocyte phase. It is not clear whether, besides degradation by the ubiquitin proteasome system, autophagy also plays a role in the elimination of sperm mitochondria [26–28]. Sutovsky and collaborators also showed that the elimination of ubiquinated sperm mitochondria could be prevented by injection of anti-ubiquitin antibodies. In parallel to mitochondria elimination, paternal mtDNA degradation probably involves a mitochondrial endonuclease that degrades mtDNA within paternal mitochondria after fertilization. However, there are reports suggesting that elimination or reduction of number of copies of paternal mtDNA molecules in mammalian sperm may have happened before fertilization [29]. As a consequence, it is nowadays largely recognized that reduction or elimination of mtDNA copies in sperm mitochondria, followed by elimination of sperm mitochondria after fertilization is the most plausible biological explanation for the strictly maternal inheritance of mammalian mtDNA [27].

Nevertheless, some striking exceptions to maternal inheritance in humans were reported. In 2002, a patient with mitochondrial myopathy, due to a deletion of 2 bp in the *MT-ND2* gene was described. The patient, born from an unaffected couple, showed mutated mtDNA in high frequency of heteroplasmy in muscle. Molecular analysis revealed that, besides heteroplasmy regarding the 2 bp deletion, the patient also showed other heteroplasmic sites, in nucleotide positions known for harboring SNPs that allowed to define mtDNA haplogroups. This indicated that he had mtDNA molecules from two different haplogroups, one inherited from the mother and the second, from the father. Haplotype analysis also allowed the conclusion that the deletion was present in the paternally derived mtDNA haplogroup, although it was not detected in the father, at least in blood. It was demonstrated in that study that the paternally derived mtDNA contributed with 90% to the mtDNA pool in skeletal muscle. The deletion probably arose de novo in early embryogenesis or, more likely, in the paternal germ line [30]. After this report, many other series of patients with mitochondrial diseases were investigated and no other paternally inherited cases of mtDNA were reported [31].

The recognized capacity of NGS in detecting DNA sequences even if they are present in very low frequencies in biological samples gave opportunity of reappraising the issue of the escape of paternal mtDNA after fertilization in humans. Studies demonstrated that low frequency heteroplasmy is a generalized phenomenon in human cells, both in normal and cancer tissues [20, 21, 32]. The comparison of heteroplasmic variants present in trios (father, mother and child) allowed the conclusion that low frequency heteroplasmic variants in children do not result from inheritance of variants present in paternal mtDNA molecules, but they raise probably due to chance, from post-zygotic changes, leading to somatic mosaicism [33].

In 2018, Luo et al. [34] reported three unrelated Chinese pedigrees in which high levels of mtDNA heteroplasmy were detected in more than one generation. Investigation of mtDNA in the three pedigrees showed evidence of biparental mtDNA transmission, and the capacity of transmitting male mtDNA seemed to be inherited as an autosomal dominant character. One of the hypothesis to explain the findings was an inherited disruption of normal cell processes related to the prevention of inheritance of paternal mtDNA. Nevertheless, these findings were recently contested by other authors that highlighted that the transmission of

NUMTs can lead to the false conclusion that paternal mtDNA is transmitted, and this would provide a more likely explanation for the recent reports of paternal transmission of mtDNA [35]. Meanwhile, while the controversy remains, it seems that paternal transmission of mtDNA must be considered as an extremely rare event and it does not change the fact that, for genetic counseling purposes, diseases that result from pathogenic variations in the mtDNA are maternally transmitted as a rule and the risk of transmission of paternal DNA is negligible [17]. Furthermore, no theoretical premises in studies about human evolution and dispersion must change to account for paternal transmission of mtDNA, because of its rarity.

10.6 Frequent mtDNA Variants Define Haplogroups Correlated to the Geographical Patterns of Dispersion of Modern Humans

The elevated variability of sequences within the mtDNA and the lack of recombination between paternal and maternal genomes allowed maternal lines in mtDNA, in all different human populations, to be transmitted as haplotype blocks. Diversity within mtDNA has been investigated all over the world by sequencing the most variable parts of the control region, the hypervariable segments I and II (HVS-I and HVS-II), often complemented by genotyping some informative SNPs from the coding region. Commonly inherited mtDNA variants have thus created stable population subgroups sharing maternal lines that can be identified according to the presence or absence of some polymorphic sequence variants. The classification of maternal lines in groups sharing some ancestral variants resulted in what we call mtDNA haplogroups [12, 36, 37].

Mitochondrial haplogroups are groups of mitochondrial sequences sharing some nucleotide variants, which indicate common ancestors. Since some variants originated in specific geographical regions and were spread with human populations' dispersal, a mitochondrial haplogroup can be a marker of geographical or even continental ancestry, on the maternal side. Most of the mtDNA of Europeans belong to one of 10 major (top level) haplogroups: H, I, J, K, R, U, T, V, W and X. A, B, C, D, E, F, G, M, N, O, P, Q, S, Z and Y are the haplogroups found originally in Asia. All L* lineages are African: L0, L1, L2, L3, L4, L5, L6. In Native American populations, only the haplogroups A, B, C, and D, which originally rose in Asia, are usually present. Each major (top level) haplogroup can be divided into many sub-haplogroups or lineages, when detailed information about mtDNA sequence is available. This has offered valuable opportunities for investigating human origins, the dispersal of human populations in the continents, its timing, and to assess genetic diversity and admixture of human populations. In addition, when coupled to Y chromosome variation studies, mtDNA allows investigation of sex-biased admixture. In other words, genetic admixture estimates based on uniparental markers, such as the

Y chromosomes and mtDNA, allow the identification of sex bias in the makeup of an admixed population. For instance, admixed populations in South America result from recent colonial admixture from European colonizers. While high frequency of Y chromosome of European ancestry indicate male-biased transfer of European DNA, mtDNA in South America comparatively reveals low frequencies of European mtDNA haplogroups. On the other hand, Native American and African mtDNA lineages are more abundant. This reveals that in South America, the European genetic contribution originated mainly from males, and that Native American and African women, and not European women, were prominent in the origin of the admixed populations [12].

MtDNA is frequently used in evolutionary studies as a genetic marker of diversity, since the rate of substitution of nucleotides is increased when compared to nuclear DNA. Remarkable variability of sequences is found mainly in the "hypervariable regions", HVS-1, HVS-2 and HVS-3, located in the D-loop region, where the rates of mutation were estimated to be the highest. Since mtDNA is present in cells in high number of copies, when compared to nuclear genes, it can be easily amplified by PCR (polymerase chain reaction). In forensic science and practice, amplification of mtDNA followed by genotyping or sequencing is possible even when samples are obtained from poorly conserved biological material, in situations when very little DNA is available or it is partially degraded. These properties of mtDNA allowed the investigation of samples obtained from ancient human remains, including bones of thousands of years of age, resulting in interesting academic outcomes to archeology and to studies of human evolution.

10.7 Frequent mtDNA Variants Correlate with Increased Susceptibility to Complex Disorders

Some mtDNA variants were clearly correlated with the origin of known maternally inherited diseases, being causative of dysfunctions with profound effects on quality of life, as already reviewed in this chapter. However, some mtDNA variants, including single nucleotide substitutions with milder effects on mitochondria functioning, can confer increased susceptibility to disease. Many mtDNA SNPs have historically segregated in haplogroups, in human evolution and migrations. Although they were often treated as "neutral" from the point of view of evolution, there are examples in which they have been found to be important in adaptation of human populations to new environments and in modulating risk of developing disease [38, 39]. Many studies correlated mitochondrial haplogroups with longevity, athletic performance, adaptation to high altitude, and risks for diabetes, Alzheimer and Parkinson diseases, some psychiatric disorders and cancer. The mtDNA haplogroups may also influence the penetrance of autosomal genetic defects or even mtDNA defects [17]. The same rare mtDNA variant can cause different degrees of severity of disease, depending on the mtDNA haplogroup on which is present. Modest but relevant differences in respiratory chain and mtDNA copy number may be present in individuals with different haplogroups. In one study, different patterns of gene expression were found in stem cells harboring different haplogroups. ROS and metabolic intermediates derived from mitochondrial metabolism act as signals that convey information between mitochondria and the nucleus. These metabolites act as substrates and co-factors for chromatin remodeling complexes, resulting in epigenetic marks that may alterate nuclear gene expression. Thus, mito-nuclear crosstalk and its link to the epigenomes may provide a way to explain why common variation in mtDNA may influence susceptibility to diseases and physiological responses to environment [40].

10.8 Somatic Variation Accumulated in the mtDNA is Related to Aging and Diseases

Accumulation of mutations in the mitochondrial genome seems to be a natural feature of aging. In a set of postulates, known in the literature as the "mitochondrial theory of aging", it was proposed that the progressive accumulation of somatic mutations in the mtDNA during lifetime leads to mitochondrial abnormalities and decline in mitochondrial function. Mitochondrial abnormalities and mtDNA mutations are instigators of multisystem degeneration and energy deficits, and one of the most important factors in this process is supposed to be the production of ROS during normal functioning of the respiratory chain taking place within mitochondria. The accumulated somatic mtDNA mutations due to ROS production can impair the function of the respiratory chain and lead to increased ROS production. As a result, more mutations are accumulated, in a vicious cycle. This cycle is believed to account for increase in oxidative damage during aging. The consequence of this cycle is loss of cellular functions, increasing energy insufficiency, cell senescence and apoptosis [4, 41, 42].

Some studies have consistently shown that mitochondrial respiration decreases with age, attributed to reduced activity in each of the four OXPHOS complexes of the mitochondrial electron transport chain (ETC). The detection of human cells deficient in COX (Cytochrome C oxidase) in aging post-mitotic tissues was the first biochemical evidence of the mitochondrial theory of aging. Loss of structure in cristae, alteration of mitochondrial morphology and dysregulation of mitochondrial metabolism are considered senescence markers, placing mitochondria as a senescence gatekeeper. Reduced mitochondrial quality and content in tissues is indeed implicated in several aging conditions such as cancer, type 2 diabetes, osteoporosis, dementia, neurologic and neurometabolic syndromes. Sustained mitochondrial dysfunction leads to activation of the caspase cascade culminating in DNA fragmentation, a characteristic of apoptosis, observed in aged tissues and in many disorders. Morphological changes in mitochondria with aging correlate with increased level of oxidative mtDNA damage, for instance, 8-oxoguanine, as well as the presence of mtDNA deletions and point mutations. Hence, mtDNA replication errors accumulating during the lifespan were proposed to be the driving force of mitochondrial metabolic failure and aging.

Aging cells show some mitochondrial properties similar to those of patients with inherited mitochondrial diseases. Not surprisingly, many of the frequent characteristics present in individuals with advanced age, such as diabetes, hearing loss, cataract, neurologic alterations and muscle weakness, are present in inherited diseases due to mtDNA pathogenic variants. This reinforces the biological connection between mutated mtDNA molecules and aging. Some mutated mtDNA variants can clonally expand to high levels in individual cells and the question remains whether this happens because abnormal copies of mtDNA are selectively replicated, but it seems that drift could also explain the findings [4].

Many studies performed with aging humans and animal models confirm the connection between age and the frequency of mtDNA mutations, which is in accordance to some of the steps of the "mitochondrial theory of aging". A remarkable advance in the field was the development, by two different groups, of mouse models expressing a defective version of the PolG mtDNA polymerase, lacking its proofreading exonuclease activity. These mice acquired mtDNA mutations at a higher rate than controls, thus being called "mutators". They showed remarkable marks of premature aging such as osteoporosis, hunched appearance, weight loss, reduced adipose tissue and muscle mass [43]. These studies confirmed that somatic mtDNA alterations contribute to aging phenotypes. In spite of this compelling evidence, still many questions remain unanswered. A lot is still required to confirm that mtDNA mutations per se are causal of the aging process or if they represent collateral findings of this process. Mitochondrial gene expression, in particular, efficiency in mitochondrial translation, is likely another important issue in aging. Damage to mitochondrial components other than DNA may be also key factors in this vicious cycle.

Modern chronic diseases are boosted by excessive food intake and sedentary lifestyle, and mitochondrial biology can be the conceptual link to explain many of the epidemiological observations in the field. Oversupply, which is the excess supply of energy substrates, mainly glucose and lipids, was correlated to biochemical mitochondrial overload, which leads to increased ROS production, mitochondrial fission, oxidative stress and these culminate in mtDNA damage, possibly increasing cellular aging, with shortening of telomeres (see Chap. 7) [3, 39].

It has already been shown that ROS is an important factor in telomere damage and mitochondria are the source of ROS. Aging in primary cells is known to be associated with a gradual increase in ROS production due to progressive mitochondrial failure and is concomitant to telomere shortening. The neutralization of ROS does not restore the mitochondrial function but inhibits telomere shortening thus establishing ROS as probably causative of telomere shortening. In addition, in human syndromes with excess of ROS production, such as some mitochondriopathies, a decrease in telomere length was observed. It is remarkable the recent increase in the number of studies aiming to connect mitochondrial processes related to senescence and aging to telomere biology [44].

Many epidemiological studies in humans demonstrate that exercise reduces the risk of several chronic diseases and contributes to increased life expectancy. Safdar and collaborators, in 2011 [45], demonstrated that endurance training was able to rescue mitochondrial biogenesis, increase mitochondria oxidative capacity, restore mitochondrial morphology, and reduce apoptosis in many tissues of the mtDNA mutator mice. Exercise attenuated the decline in mtDNA copy number and decreased the frequency of point mutations in the mtDNA, which appeared as related to the progeroid phenotype in the mutator mice. Endurance training also contributed to mitigate apoptosis, and the premature mortality was also prevented. They hypothesized that exercise may impose selective mitochondrial biogenesis of healthy mitochondria via modulation of mitochondrial dynamics, by promotion of fusion and fission and by destruction of mitochondria carrying high levels of mutated mtDNA. Their findings support exercise as an approach to improving systemic mitochondrial dysfunction caused by aging or diseases.

Exercise has been also shown to positively affect the brain and reverse age-related brain atrophy. Exercise increases whole-body oxygen consumption and accelerates mitochondrial energy production. Increased energy demand engages adaptive signaling pathways that increase mitochondrial content and optimize their function via mitochondrial biogenesis, inducing the expression of genes that restrict inflammation and may therefore counteract pro-aging mitochondrial signaling. There is also evidence that exercise stimulates mitochondrial biogenesis in the brain. In accordance, sedentary behavior is a known major risk for Alzheimer's disease. This might be explained by the fact that physical inactivity promotes metabolic stress, possibly through disruption of mitochondrial dynamics and accumulation of mtDNA damage [3].

It can be predicted that advances in understanding mitochondrial biology will soon help to develop strategies to prevent the adverse effects of aging and to treat mitochondrial diseases.

10.9 mtDNA Disorders

Current treatment strategies for conditions related to mitochondrial dysfunction are limited, since they address some of the symptoms but do not mitigate the reduced mitochondrial oxidative capacity in aged tissues. The research of novel strategies to mitigate mitochondrial dysfunction is desirable to improve the quality of life of aging subjects or of the ones affected by mitochondrial disorders. Some studies have attempted to address mitochondrial diseases focusing on different strategies of treatment: promotion of increase the oxidizing capacity of mitochondria; administration of lacking substances or adding substances that increase the energy capacity of mitochondria; reduction of the quantity of mutated mtDNA and induction of mitochondrial biogenesis. For instance, vitamins and co-factors, such as vitamins C and E, coenzyme Q10 and folic acid were tried. Some of these substances provided some beneficial effects, depending on the type of disease. However, there is remarkable interest in academic research related to the fourth strategy, the stimulation of mitochondrial biogenesis as a consequence of exercise, as mentioned in the previous paragraphs.

The most effective strategy presently available is the prevention through appropriate genetic counseling. There are empiric recurrence risks available in the case of transmission of the most common , if they are in homoplasmy. However, whenever the causative variants are in heteroplasmic state, the genetic bottleneck is a barrier to the precise prediction of disease risk in the following generations.

Pre-implantation genetic diagnosis (PGD) was shown to be of utility to some women with pathogenic variants in heteroplasmy. Embryos obtained after *in vitro* fertilization can be genetically analyzed, through biopsy of one or few blastomeres, and those embryos which are mutation-free or show the lowest mutation levels can be selected to be transferred to the uterus. Such strategy can minimize the probability of severe disease, but it may not be able to eliminate completely the risk of disease [46].

Some advances were also made in the field of pro-nuclear transfer with the aim of correcting mitochondrial disease. This involves the transfer of nuclear DNA from the donor zygote (obtained from a donor couple in which the mother has mtDNA disease) to an enucleated recipient zygote, by means of fusion. The zygote retains nuclear DNA from the parents, but the mtDNA, with wild-type sequence, comes from the recipient zygote. Other groups used a similar technique but they utilized spindle transfer, instead of pro-nuclear transfer, with equivalent results. These advances are far from being a routine in the clinical setting and they create debate because of potential long-term effects and ethical issues. Legalization of these procedures is not world spread and will be an ongoing debate [19, 46].

In parallel to investigations aiming at treating genetic disorders caused by variants in nuclear DNA, there is hope that, in the next years, research in gene therapy and genomic edition will bring some relief to carriers of mitochondrial dysfunctions caused by mtDNA alterations.

10.10 mtDNA is Related to Inflammation and Immunity

In mammalian cells, mitochondria contribute to immune and inflammation processes in different ways. Following mitochondrial damage due to oxidative stress, circular mtDNA can leak out into the cytoplasm. As a result, the NLRP3 (leucine-rich repeat (LRR)-containing proteins (NLR) family member 3) inflammasome is triggered by mtDNA outside the mitochondria. Mitochondria also act in immune signaling affecting anti-viral responses. Third, mtDNA can also leak into the systemic circulation where it is recognized by TLR9 (Toll-like receptor 9) and this leads to tissue lesion in heart, vascular system and may cause neurological degenerative states. Circulating cell-free mtDNA (ccf-mtDNA) and other molecules that are free in the blood are thus putative markers of early stages of diseases related to mitochondrial stress. Some studies suggested that ccf-mtDNA increases with age, contributing to the overall trend of increasing inflammation that happens with age [3]. There is an emerging notion that mitochondria is related to factors that can trigger inflammatory and pathological processes that underlie many common chronic diseases that increase with age, such as diabetes.

10.11 mtDNA and Epigenetics

Epigenetics is an important layer of information on DNA sequences and is a key factor for establishing profiles of gene expression. DNA can be epigenetically modified via methylation of citosines, a process that frequently leads to transcriptional silencing of genes if it occurs near promoter regions. Histones, key proteins in the assembly of nuclear chromatin, can also be epigenetically altered by post-translational modification such as acetylation, phosphorylation, methylation, sumoylation and ubiquitination of their N-terminal tails. Non-coding RNAs also play key roles in the definition of epigenetic landscapes, regulating gene expression. Histone modifications, DNA methylation and non-coding RNA expression are epigenetic modifications that explain differential nuclear gene expression in different cells types and in different stages of development, because they induce changes in chromatin states that affect initiation of transcription [47]. It has only recently been recognized that gene expression in mitochondria may also be regulated via epigenetic mechanisms, as it happens to nuclear gene expression. In parallel to epigenetic processes that regulate gene expression in nucleus, DNA methylation, non-coding RNAs and post-translational modification of proteins associated to the nucleoid, were identified within mitochondria and they are probably linked to the regulation of gene expression, although the matter is still the subject of ongoing debate [38].

The mechanisms of methylation and demethylation in mtDNA are not clearly understood. Several studies identified mtDNA methylation in cell lines and tissue samples, including of human origins. Methylation was reported in mtDNA many decades ago and it is a matter of controversy ever since. Some methodological issues complicate the estimates of overall methylation in mtDNA, such as contamination with NUMTs. The first studies focused on methylation of CpG sites, with heterogeneous and conflicting results. In the nucleus, cytosine methylation frequently occurs within CpG nucleotides clustered in CpG islands (see Chap. 4), but it may also occur in other sites. However, CpG islands are absent in mtDNA and methylation occurs within dispersed CpGs sites. Besides, non-CpG methylation, such as CpC, CpA and CpT, and adenine methylation were also observed. The studies altogether point that methylation indeed occurs within mtDNA. The endosymbiotic theory of mitochondrial origin combined with the findings of abundant adenine methylation in mtDNA suggest that adenine methylation in mitochondria may be more relevant than cytosine methylation, as seen in nuclear DNA. A mitochondrial localized DNA methyltransferase 1 (mtDNMT1) was identified in 2011, raising the debate on the role of methylation in gene regulation within mitochondria.

The mitochondrial D-loop is one of the most important regions to the expression of the mtDNA due to its role in controlling transcription and replication. Differential methylation within the D-loop, especially in cytosine nucleotides, has been described in many studies, but its precise function is unknown; it is tempting to speculate that different methylation profiles in the D-loop region would be related to mtDNA gene expression. Apart from D-loop, gene bodies also have regions in which methylation may have an effect on gene expression. Changes in methylation of mitochondrial genes were shown to correlate with changes in gene expression [38].

With the ongoing recognition that epigenetic modifications may play a role in mtDNA expression, the role of several factors on levels of mtDNA methylation was investigated. Several external factors such as air pollutants, smoking, diet and drugs were demonstrated as affecting mtDNA methylation. Some air pollutants and smoking were associated to D-loop and gene methylation in mtDNA, with important correlations to human diseases. In pigs, maternal diet was shown to alter mtDNA methylation levels in newborns, affecting their OXPHOS capacity. Differential mtDNA methyation was also correlated to many frequent human diseases such as Alzheimer, Parkinson, cancer and metabolic disorders, including obesity [38, 47].

In mitochondria, histone proteins are absent but modifications of nucleoid proteins were shown to play a role in the regulation of gene expression. Many proteins localized within mitochondria contain potential acetylation sites. TFAM, the main structural component of mitochondrial nucleoids, is a protein that promotes replication, transcription and general maintenance of mtDNA. TFAM can be modified by acetylation, glycosylation and phosphorylation. Acetylation and phosphorylation alter the affinity of TFAM to DNA, thus affecting mtDNA compaction; mtDNA compaction, as consequence, probably affects mtDNA replication and transcription. Furthermore, it has been shown that levels of TFAM occupancy in mtDNA affect the access of DNMTs to methylate DNA, showing that TFAM plays a role in the pattern of mtDNA methylation.

Acetylation and phosphorylation sites were identified in other nucleoid-associated proteins, including mtSSB and DNA polG, but their role in the regulation of gene expression is unknown.

Different classes of non-coding RNAs (ncRNAs) involved in the epigenetic regulation of gene expression in mitochondria are known, but it is not clear in some cases if they are transcribed inside the mitochondria or if they are derived from NUMTs. Several long non-coding RNAs (lncRNA) encoded by the mtDNA have been identified and there is evidence that they participate in the regulation of mitochondrial gene expression. Some of these lncRNAs are transported into the nucleus and act as retrograde signaling molecules, and they are believed to act in mito-nuclear crosstalk. Moreover, significant changes in their levels were observed in cancer, suggesting they function in cell cycle progression. There are also lncRNA molecules encoded by nuclear genes and transported into the mitochondria, where they regulate mitochondrial processes of metabolism and apoptosis.

MicroRNAs were also identified within mitochondria and they were termed mitochondrial microRNAs (mitomiRs). They are short (17–25 bp) single-stranded RNA molecules that are transcribed in nucleus from nuclear templates and are transported to mitochondria, but some of them are transcribed using mtDNA as template. They regulate expression of nuclear-encoded and mtDNA-encoded proteins. There is evidence that they can enhance or repress gene expression, at transcriptional and translational levels, modulating metabolic activities [38, 47].

There is compelling evidence that there are many pathways allowing exchange of information between mitochondria and nucleus, and this exchange may affect gene expression in the nucleus. The signals that convey information between mitochondria and nucleus are ROS and other metabolic intermediates from mitochondrial metabolism. Some of these metabolites are required substrates and co-factors in chromatin remodeling complexes, influencing post-translation histone tail modifications by histone acetylases, histone deacetylases, for instance, or leading to DNA modifications via DNA methyltransferases or demethylases. The resulting biochemical changes influence the epigenetic state of nuclear genes, resulting in changes in gene expression. It can be concluded that epigenetic communication between nuclear and mitochondrial genomes occurs at multiple levels, ensuring a coordinated gene expression between these two different genetic compartments. Metabolic changes stimulated, for example, by environment factors, such as diet or physical activity, alter the relative abundances of various metabolites, directly affecting the epigenetic machinery, both in mitochondria and nucleus [3, 47].

Acknowledgements I am greatly indebted to Dr. Lilian Kimura and Ms. Vinícius Magalhães Borges for formatting and checking references and figure.

References

1. Alberts B, Hopkin K, Johnson AD, Morgan D, Raff M, Roberts K, Walter P. Essential cell Biology, 5th International Student Edition. 2018. Wiley.
2. Chinnery PF, Hudson G. Mitochondrial genetics. Br Med Bull. 2013;106(1):135–59. https://doi.org/10.1093/bmb/ldt017. Epub 2013 May 22. PMID: 23704099; PMCID: PMC3675899.
3. Picard M, Wallace DC, Burelle Y. The rise of mitochondria in medicine. Mitochondrion. 2016;30:105–16.
4. Taylor RW, Turnbull DM. MtDNA mutations in human disease. Nat Rev Genet. 2005 May;6(5):389–402.
5. Margulis L. The origin of plant and animal cells. Am Sci. 1971;59(2):230–5. PMID:5170543.
6. Gray MW. Rickettsia, typhus and the mitochondrial connection. Nature. 1998;396(6707):109–10. https://doi.org/10.1038/24030. PMID:9823885.
7. Gray MW, Burger G, Lang BF. The origin and early evolution of mitochondria. Genome Biol. 2001;2(6):REVIEWS1018. https://doi.org/10.1186/gb-2001-2-6-reviews1018.
8. Thrash JC, Boyd A, Huggett MJ, et al. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci Rep. 2011;1:13. https://doi.org/10.1038/srep00013.

9. Bernt M, Braband A, Schierwater B, Stadler PF. Genetic aspects of mitochondrial genome evolution. Mol Phylogenet Evol. 2013;69(2):328–38. https://doi.org/10.1016/j.ympev.2012.10.020. Epub 2012 Nov 7. PMID:23142697.
10. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. Sequence and organization of the human mitochondrial genome. Nature. 1981 Apr 9;290(5806):457–65.
11. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mtDNA. Nat Genet. 1999 Oct;23(2):147.
12. Jobling M, Hollox E, Kivisild T, Tyler-Smith C. Human evolutionary genetics. 2nd ed. New York: Garland Science; 2013 June 25.
13. Wallace DC, Zheng XX, Lott MT, Shoffner JM, Hodge JA, Kelley RI, Epstein CM, Hopkins LC. Familial mitochondrial encephalomyopathy (MERRF): genetic, pathophysiological, and biochemical characterization of a mtDNA disease. Cell. 1988;55(4):601–10.
14. DiMauro S, Schon EA. Mitochondrial respiratory-chain diseases. N Engl J Med. 2003;348(26):2656–68.
15. Alston CL, Rocha MC, Lax NZ, Turnbull DM, Taylor RW. The genetics and pathology of mitochondrial disease. J Pathol. 2017;241(2):236–250. https://doi.org/10.1002/path.4809. Epub 2016 Nov 2. PMID: 27659608; PMCID:PMC5215404.
16. Prezant TR, Agapian JV, Bohlman MC, Bu X, Oztas S, Qiu WQ, Arnos KS, Cortopassi GA, Jaber L, Rotter JI, et al. Mitochondrial ribosomal RNA mutation associated with both antibiotic-induced and non-syndromic deafness. Nat Genet. 1993;4(3):289–94. https://doi.org/10.1038/ng0793-289. PMID:7689389.
17. Estivill X, Govea N, Barceló E, Badenas C, Romero E, Moral L, Scozzri R, D'Urbano L, Zeviani M, Torroni A. Familial progressive sensorineural deafness is mainly due to the mtDNA A1555G mutation and is enhanced by treatment of aminoglycosides. Am J Hum Genet. 1998;62(1):27–35. https://doi.org/10.1086/301676. PMID:9490575; PMCID:PMC1376822.
18. Wei W, Chinnery PF. Inheritance of mitochondrial DNA in humans: implications for rare and common diseases. J Inter Med. 2020;287(6):634–44.
19. Craven L, Alston CL, Taylor RW, Turnbull DM. Recent advances in mitochondrial disease. Annu Rev Genomics Hum Genet. 2017a Aug 31;18:257–75.
20. Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high-throughput sequencing of complete human mtDNA genomes. Am J Hum Genet. 2010 Aug 13;87(2):237–49.
21. Sosa MX, Sivakumar IK, Maragh S, Veeramachaneni V, Hariharan R, Parulekar M, Fredrikson KM, Harkins TT, Lin J, Feldman AB, Tata P, Ehret GB, Chakravarti A. Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol. 2012;8(10):e1002737.
22. Ankel-Simons F, Cummins JM. Misconceptions about mitochondria and mammalian fertilization: implications for theories on human evolution. Proc Natl Acad Sci U S A. 1996 Nov 26;93(24):13859–63.
23. Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G. Ubiquitin tag for sperm mitochondria. Nature. 1999;402(6760):371–2.
24. Al Rawi S, Louvet-Vallée S, Djeddi A, Sachse M, Culetto E, Hajjar C, Boyd L, Legouis R, Galy V. Postfertilization autophagy of sperm organelles prevents paternal mitochondrial DNA transmission. Science. 2011;334(6059):1144–7. https://doi.org/10.1126/science.1211878. Epub 2011 Oct 27. PMID:22033522.
25. Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G. Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos. Biol Reprod. 2000 Aug;63(2):582–90.
26. Luo SM, Schatten H, Sun QY. Sperm mitochondria in reproduction: good or bad and where do they go? J Genet Genomics. 2013 Nov 20;40(11):549–56.

27. Song WH, Ballard JW, Yi YJ, Sutovsky P. Regulation of mitochondrial genome inheritance by autophagy and ubiquitin-proteasome system: implications for health, fitness, and fertility. Biomed Res Int. 2014;2014:981867.
28. Zhou Q, Li H, Li H, Nakagawa A, Lin JL, Lee ES, Harry BL, Skeen-Gaar RR, Suehiro Y, William D, Mitani S, Yuan HS, Kang BH, Xue D. Mitochondrial endonuclease G mediates breakdown of paternal mitochondria upon fertilization. Science. 2016;353(6297):394–9. https://doi.org/10.1126/science.aaf4777. Epub 2016 Jun 23. PMID:27338704; PMCID:PMC5469823.
29. Schwartz M, Vissing J. Paternal inheritance of mtDNA. N Engl J Med. 2002 Aug 22;347(8):576–80.
30. Schwartz M, Vissing J. New patterns of inheritance in mitochondrial disease. Biochem Biophys Res Commun. 2003 Oct 17;310(2):247–51.
31. He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, Diaz LA Jr, Kinzler KW, Vogelstein B, Papadopoulos N. Heteroplasmic mtDNA mutations in normal and tumour cells. Nature. 2010 Mar 25;464(7288):610–4.
32. Pyle A, Hudson G, Wilson IJ, Coxhead J, Smertenko T, Herbert M, Santibanez-Koref M, Chinnery PF. Extreme-depth re-sequencing of MtDNA finds no evidence of paternal transmission in humans. PLoS Genet. 2015 May 14;11(5):e1005040.
33. Luo S, Valencia CA, Zhang J, Lee NC, Slone J, Gui B, Wang X, Li Z, Dell S, Brown J, Chen SM, Chien YH, Hwu WL, Fan PC, Wong LJ, Atwal PS, Huang T. Biparental inheritance of MtDNA in humans. Proc Natl Acad Sci U S A. 2018 Dec 18;115(51):13039–44.
34. Wei W, Pagnamenta AT, Gleadall N, Sanchis-Juan A, Stephens J, Broxholme J, Tuna S, Odhams CA, Genomics England Research Consortium, NIHR BioResource, Fratter C, Turro E, Caulfield MJ, Taylor JC, Rahman S, Chinnery PF. Nuclear-mitochondrial DNA segments resemble paternally inherited mitochondrial DNA in humans. Nat Commun. 2020;11(1):1740. https://doi.org/10.1038/s41467-020-15336-3. Erratum in:Nat Commun. 2020;11(1):3741. PMID: 32269217; PMCID: PMC7142097.
35. MITOMAP https://www.mitomap.org/MITOMAP
36. Wallace DC. Mitochondrial DNA mutations in disease and aging. Environ Mol Mutagen. 2010 Jun;51(5):440-50. https://doi.org/10.1002/em.20586. PMID: 20544884.
37. Brandon MC, Lott MT, Nguyen KC, Spolim S, Navathe SB, Baldi P, Wallace DC. MITOMAP: a human mitochondrial genome database--2004 update. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D611–3.
38. Mposhi A, Van der Wijst MG, Faber KN, Rots MG. Regulation of mitochondrial gene expression, the epigenetic enigma. Front Biosci (Landmark Ed). 2017 Mar 1;22:1099–113.
39. Wallace DC. Genetics: Mitochondrial DNA in evolution and disease. Nature. 2016;535(7613):498–500. https://doi.org/10.1038/nature18902.
40. Bonawitz ND, Shadel GS. Rethinking the mitochondrial theory of aging: the role of mitochondrial gene expression in lifespan determination. Cell Cycle. 2007;6(13):1574–8. https://doi.org/10.4161/cc.6.13.4457. Epub 2007. PMID:17603300.
41. Larsson NG. Somatic mitochondrial DNA mutations in mammalian aging. Annu Rev Biochem. 2010;79:683–706. https://doi.org/10.1146/annurev-biochem-060408-093701. PMID:20350166.
42. Trifunovic A, Wredenberg A, Falkenberg M, Spelbrink JN, Rovio AT, Bruder CE, Bohlooly-Y M, Gidlöf S, Oldfors A, Wibom R, Törnell J, Jacobs HT, Larsson NG. Premature ageing in mice expressing defective mtDNA polymerase. Nature. 2004 May 27;429(6990):417–23.
43. Sahin E, Colla S, Liesa M, Moslehi J, Müller FL, Guo M, Cooper M, Kotton D, Fabian AJ, Walkey C, Maser RS, Tonon G, Foerster F, Xiong R, Wang YA, Shukla SA, Jaskelioff M, Martin ES, Heffernan TP, Protopopov A, Ivanova E, Mahoney JE, Kost-Alimova M, Perry SR, Bronson R, Liao R, Mulligan R, Shirihai OS, Chin L, DePinho RA. Telomere dysfunction induces metabolic and mitochondrial compromise. Nature. 2011;470(7334):359–65. https://doi.org/10.1038/nature09787. Epub 2011 Feb 9. Erratum in:Nature. 2011 Jul 14;475(7355):254. PMID:21307849; PMCID:PMC3741661.

44. Safdar A, Bourgeois JM, Ogborn DI, Little JP, Hettinga BP, Akhtar M, Thompson JE, Melov S, Mocellin NJ, Kujoth GC, Prolla TA, Tarnopolsky MA. Endurance exercise rescues progeroid aging and induces systemic mitochondrial rejuvenation in mtDNA mutator mice. Proc Natl Acad Sci U S A. 2011 Mar 8;108(10):4135–40.
45. Craven L, Tang MX, Gorman GS, De Sutter P, Heindryckx B. Novel reproductive technologies to prevent mitochondrial disease. Hum Reprod Update. 2017b Sep 1;23(5):501–19.
46. Vaiserman AM, Koliada AK, Jirtle RL. Non-genomic transmission of longevity between generations: potential mechanisms and evidence across species. Epigenetics Chromatin. 2017;10(1):38. https://doi.org/10.1186/s13072-017-0145-1. PMID:28750655; PMCID:PMC5531095.
47. Sharma N, Pasala MS, Prakash A. MtDNA: epigenetics and environment. Environ Mol Mutagen. 2019 Oct;60(8):668–82.

Chapter 11
Population Variation of the Human Genome

Fabrício R. Santos, Thomaz Pinotti, and Ricardo Fujita

11.1 What Is Genomic Variation?

The human genome is a dynamic storage of information whose inherited DNA changes are created by mutation and recombination (chromosomal segregation and crossing-over) during meiosis to produce gametes. These DNA changes or genotypic variants are accumulated and reshaped in subsequent generations of populations under influence of random (drift) and deterministic (selection) evolutionary mechanisms. However, only genotypic variants affecting phenotypes are potentially related to health and clinical conditions and we still know relatively little about the direct relationship between genotypes and phenotypes. Moreover, the distinction between neutral (not expressed at phenotype level) and functional variation is not straightforward, as well as their association to environmental factors. A major cause of this uncertainty today is related to our poor knowledge about gene expression,

Genes cited in this chapter: *ACE* (Angiotensin I-Converting Enzyme); *ACE2* (Angiotensin-Converting Enzyme 2); *ASPM* (Assembly Factor for Spindle Microtubules); *CCR5* (C-C motif chemokine Receptor 5); *FOXP2* (Forkhead box P2); *FURIN* (Furin paired basic amino acid cleaving enzyme); *IL12RB1* (Interleukin 12 Receptor subunit Beta 1); *LCT* (Lactase); *MCPH1* (Microcephalin 1); *PDE10A* (Phosphodiesterase 10A); *TMPRSS2* (Transmembrane serine protease 2); *TYK2* (Tyrosine Kinase 2).

F. R. Santos (✉)
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
e-mail: fsantos@icb.ufmg.br

T. Pinotti
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

University of Copenhagen, Copenhagen, Denmark

R. Fujita
Universidad San Martin de Porres, Lima, Peru

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9_11

epistasis and epigenetics throughout all developmental stages from zygote to adulthood, when phenotypes are differentially expressed in human cells, tissues, structures and organs. Thus, the association between genes, genotypes (variants) and their phenotypes is also currently investigated by comparative and ontogenetic methods of the Evolutionary Developmental Biology field, or simply Evo-Devo.

The technological breakthroughs in genome sequencing in the last three decades allowed the description of genetic variation in an unprecedented pace. In 1994, a pioneering work by Cavalli-Sforza and colleagues described a "high resolution" tree of human populations using a "large number of loci": 30 in total [1]. In 2020, a study analyzing the immortalized cell lines from the same samples as Cavalli-Sforza's work described a total of 76.1 million variants [2]. This several-fold difference in 26 years was made possible by large collaborating efforts, as the sequencing of the Human Genome [3], the SNP Consortium [4] and the HapMap project [5], that collectively described around ten million variants. The 1000 Genomes Project (1KGP) was responsible for another great leap forward that, despite the name, ended up with 2504 individuals from 26 populations [6–8]. The 1KGP described 88 million variants, combining low-coverage shotgun sequencing, exome capture and microarray techniques, and created a free and easily available database for studying human variation, allowing to demonstrate many genotype-phenotype associations confidently [9]. As an example, there are currently 6721 clinically relevant phenotypes with known DNA variants distributed in 4316 human genes reported only in one database, the Online Mendelian Inheritance in Man (omim.org—accessed in August 27, 2020).

The databases of genetic variants allowed the comparison of human genomes revealing that variation occurs mostly as Single Nucleotide Polymorphisms (SNPs or SNVs), small insertions and deletions (indels, <50 bp), and structural variations (SVs). In general, SVs comprise different classes of polymorphisms: mobile element insertions (transposons), copy number variants (CNVs), indels, duplications and inversions of different sizes, as well as inter- and intrachromosomal translocations and other complex rearrangements. Furthermore, SVs are also known to affect several different phenotypes and are associated to many diseases and syndromes [10]. A typical human genome is estimated to contain 2100—2500 SVs: 1000 large deletions, 160 CNVs, 915 Alu insertions, 128 L1 insertions, 51 complex rearrangements, four nuclear mitochondrial DNA (NUMT) insertions and 10 inversions [8]. However, 99.9% of the genomic variation is composed of SNPs, i.e. ~5 million variable nucleotide sites in a typical human genome. Even though SNPs are the most common type of variation, SVs affect larger segments of the genome (˜20 million base pairs) and are much more difficult to analyze with current technology [10].

Despite the large genomic diversity found in modern humans, it pales in comparison to other great apes. A preliminary study [11] sequenced only 24 common chimpanzees (*Pan troglodytes*) to describe more than 27 million SNPs, while sequencing 1092 humans added up 38 million SNPs [7]. Again, this is an important reminder of how recent (in an evolutionary timescale in comparison to other apes) is our species and its global expansion out of Africa (see below).

11.2 An Evolutionary Perspective on Human Diversity

Every one of the 8 billion humans alive today carry two chromosome sets in their diploid genome, but none of those chromosomes are identical and that makes each individual unique and unrepeatable in history. The sum of all these genomic differences is what is called human genetic diversity. While this seems, and indeed it is, dazzling, this heritable variation also bears witness to the natural history of modern humans, and therefore is properly understood and studied under an evolutionary framework. Although modern human populations result from a long evolutionary history since anatomically modern *Homo sapiens* arose in Africa at least 200,000 years ago—ya [12], the history of our genome is much older than this. In fact, a large part of the human genome still retains a vast legacy of our primate ancestry. For example, a major part of our genomic architecture is shared practically unchanged with the two chimpanzee species *Pan paniscus* (bonobo) and *P. troglodytes* (common chimpanzee), our closest extant relatives. Indeed, there is almost a complete synteny of genes between human and chimpanzee chromosomes, and about 83% of the assembled genomes of all great apes (humans, chimpanzees, gorillas and orangutans) are found in multiple sequence alignments of syntenic blocks [13].

Besides explaining how genes are organized, our recent common ancestry with chimpanzees and gorillas explains also many alleles in many loci influenced by balancing selection, particularly in the Major Histocompatibility Complex (*MHC*) genes (see below), which are named trans-species polymorphisms [14]. Balancing natural selection favors multiple alleles in *MHC* genes, an adaptive diversity that increases immune response to an immense number of antigens constantly affecting primate populations over 80 million years of evolution of this mammal order. It also explains many identical *MHC* alleles conserved between humans, bonobos and common chimpanzees, even though they share a common ancestor at about six million ya [15] or 240,000 human generations (~25 years per generation). Otherwise, variants found in genes known to be functionally different between *Homo* and *Pan* were suggested to explain some exclusive characteristics of the modern humans. For example, the human *FOXP2* gene was initially associated to inner speech and speech fluency, and the human *ASPM* and *MCPH1* genes were correlated with enlarged brain size, even though no conclusive signs of positive selection were found for these genes so far [16, 17]. Other genomic comparisons included also ancient DNA data of extinct hominins like Neanderthals and Denisovans (see below), which generated a comprehensive list of potential SNPs associated to distinctive characteristics of modern humans [18].

Taking into account the dynamics of intergenerational accumulation of variation in the human genome, the last 8000 generations or 200,000 years of *Homo sapiens* history since its African origin can be roughly divided in two very distinctive periods: (1) the indigenous settlement of the world, an (2) the formation of the cosmopolitan civilization.

The Indigenous World was colonized by descendants of the first African populations of *Homo sapiens* that were already established in all continental landmasses at about 18,000 ya, when America was finally settled [19]. By this time, all population dispersals mainly took place by humans travelling on foot or eventually using small rudimentary boats. In this long period of our history, populations accumulated most of our neutral and adaptive variation according to the different inhabited environments, which explains also some of the few characteristic phenotypes commonly found in indigenous peoples of each continent. The first continental settlers were all hunter-gatherers, like some currently isolated indigenous tribes in the Amazon Forest, Andaman Islands (Indian Ocean), Africa and Southeast Asia are today, including few contemporaneous societies with no contact with modern urban society (https://www.survivalinternational.org). This period was accompanied by many cultural revolutions in human societies, with the first symbolic representations in art starting at about 100,000 ya in Africa [20].

Undoubtedly, however, the most dramatic revolution of all was probably the sedentarism and urbanization promoted by farming and pastoralism starting at about 12,000 ya, when the warmer and more humid climate of the Holocene began. With the transition from gathering to agriculture and the concomitant domestication of plants and animals [21, 22], many hunter-gatherers were dominated by or became farmers and/or pastoralists in most parts of the world, a scenario also supported by linguistic evidence [23]. After the initial settlement of continents, they have also expanded to previously uninhabited lands with the help of important technical advances in navigation, occupying for example many remote oceanic islands in Micronesia and Polynesia. As a whole, this period was marked by a long period of dispersal towards new lands and short-range intermarriages, i.e., parents were usually born in nearby places. It means that most of gene flow occurred regionally, in an intracontinental level [24, 25].

Many recent advances on the studies of human variation come from the analyses of ancient genomes. The technology of extracting and sequencing DNA of ancient human remains, often called archaeogenetics or just ancient DNA [26], has made clear how big was the contribution of agriculturalist populations to the present-day gene pool. Modern Europeans, for example, descend almost entirely from farming and pastoralist populations from the Near East and the Eurasian Steppe [27–29], with almost all of the paternal lineages in Iberia today being of Steppe origin [30]. A similar scenario seems to have occurred in Southeast Asia, where past farming populations also contributed disproportionately to present-day populations [31]. In short, in this initial period, several indigenous (native) societies flourished in Africa, Europe, Asia, Oceania and Americas, with a small or negligible contact with populations from other continents until the XV century (Fig. 11.1).

Overall, a very important insight from ancient genomes is that every population—not rarely treated as a static unit in human genetics research—is, in fact, the result of a heavily admixed story of past indigenous groups, that oftentimes were also genetically differentiated [32–34]. In other words, the indigenous past is much more complex and has to be recognized by every human geneticist interested in history or diseases [35, 36].

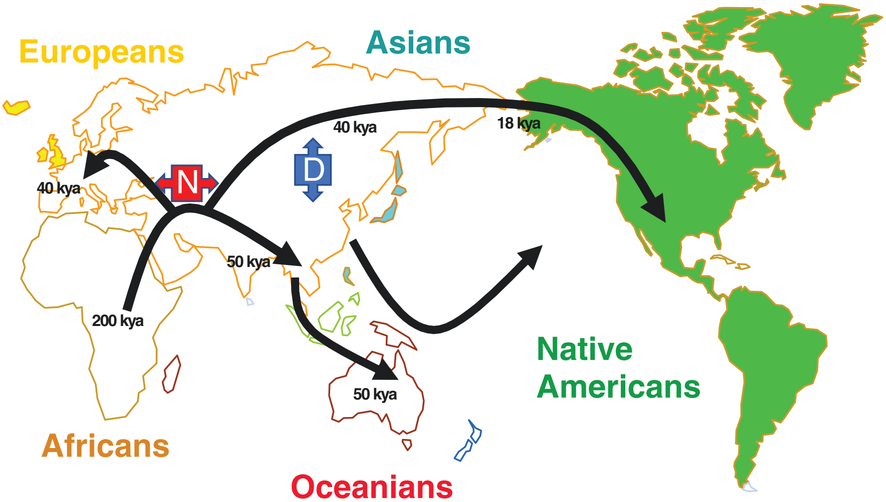


Fig. 11.1 The indigenous peopling of the world. Anatomically modern humans originated in Africa 200,000 ya and peopled all other continents during the last 70,000 years (black arrows), acquiring also some genomic variation from our close extinct cousins: Neanderthals (N) and Denisovans (D)

Nevertheless, prior knowledge of ancient and recent population history is of paramount importance for epidemiological issues related to human variation, and to be considered in many genetic analyses, including genome-wide association studies (GWAS), as hidden patterns of shared ancestry are known to be confounding factors that can both drive false positives or lower the statistical significance of true positives [37, 38].

In the subsequent period of human history there was a complete rearrangement of the human societies towards a Cosmopolitan World, when the first intercontinental human assemblages were formed, enabled by the transcontinental navigations beginning in the XV century, particularly on the way to the New World. Since then, human societies worldwide have been experiencing an increasing admixture of gene pools previously restricted to indigenous groups of each continent, further amplified by wars, migration and the cultural revolutions of the XIX and XX centuries that allowed the current globalization, towards a "single" Cosmopolitan Civilization. This period was (and still is) marked by intercontinental intermarriages, with parents coming from birthplaces increasingly farther away, generating children with admixed genomes derived from many ancestors originally belonging to different indigenous populations relatively isolated in their continents until the XV century.

Although this is beyond the topic of this chapter, it must be noted that the first intercontinental contacts were far from peaceful, and were marked by wars of extermination, enslavement, rape, forced migrations, epidemics and destruction of traditional modes of subsistence, languages and cultural practices. This colonial dominance can be seen also in the genetic data, with a strong sex bias towards

European paternal lineages in contrast with African and Native American lineages throughout all of the American continent [39–42]. In addition, there is evidence of a huge loss of Native American genetic diversity in contemporary Americas compared with the much higher diversity found in pre-Columbian populations [43, 44]. This is further supported by historical and archaeological evidence showing that all indigenous populations in the Americas were severely reduced during European colonization [45].

In conclusion, these two periods of human evolution were shaped by remarkably different cultural stages related to the connectivity of human societies, which explain the worldwide human genome variation observed in the XXI century.

11.3 Archaic Hominin Introgression as an Unexpected Source of Variation

The aforementioned insights brought by ancient DNA on the diversity and complexity of the history of populations in the Indigenous period drew a more detailed view of our past. Even more surprising was the discovery that our closest extinct relatives, the Neanderthals, contributed to a small part of the genome of all indigenous peoples outside Sub-Saharan Africa (Fig. 11.2, [46, 47]), likely admixing with anatomically modern humans on their range expansion out of Africa (Fig. 11.1). This biological phenomenon is named introgression and is observed after past hybridization events that allowed some DNA variants from one species to be incorporated in the genomic background of another species. It happens also in natural and captive populations of many vertebrates, such as the observed in the common chimpanzee that shows an introgression sign derived from the bonobo [48].

More surprisingly, Neanderthals were not the only hominin species contributing to our genome. Another archaic population named ‘Denisovans’, known only from sequencing data from few bone pieces, has also admixed in a distant past with ancient populations (Fig. 11.1) giving rise to indigenous East Asians, Southeast Asians, Native Americans and Polynesians, but also contributing to around 4% of the genetic make-up of indigenous Australians and Melanesians [49, 50]. However, recent estimates place the Denisovan contribution in Melanesians to similar levels as Neanderthals in non-Africans (Fig. 11.2 [2]). While those discoveries were only made when the genetic data of ancient individuals were available, it is also possible to infer old admixture events based on sequenced variants of modern humans without the use of an ancient DNA reference. For example, there is relative confidence of one or multiple events of archaic hominin introgression taking place also in Sub-Saharan Africa [51, 52], even though no ancient hominin from Africa has been sequenced so far. However, due to current inherent difficulty of accessing the specific genomic variants involved in this inferred admixture, they will not be further discussed here.

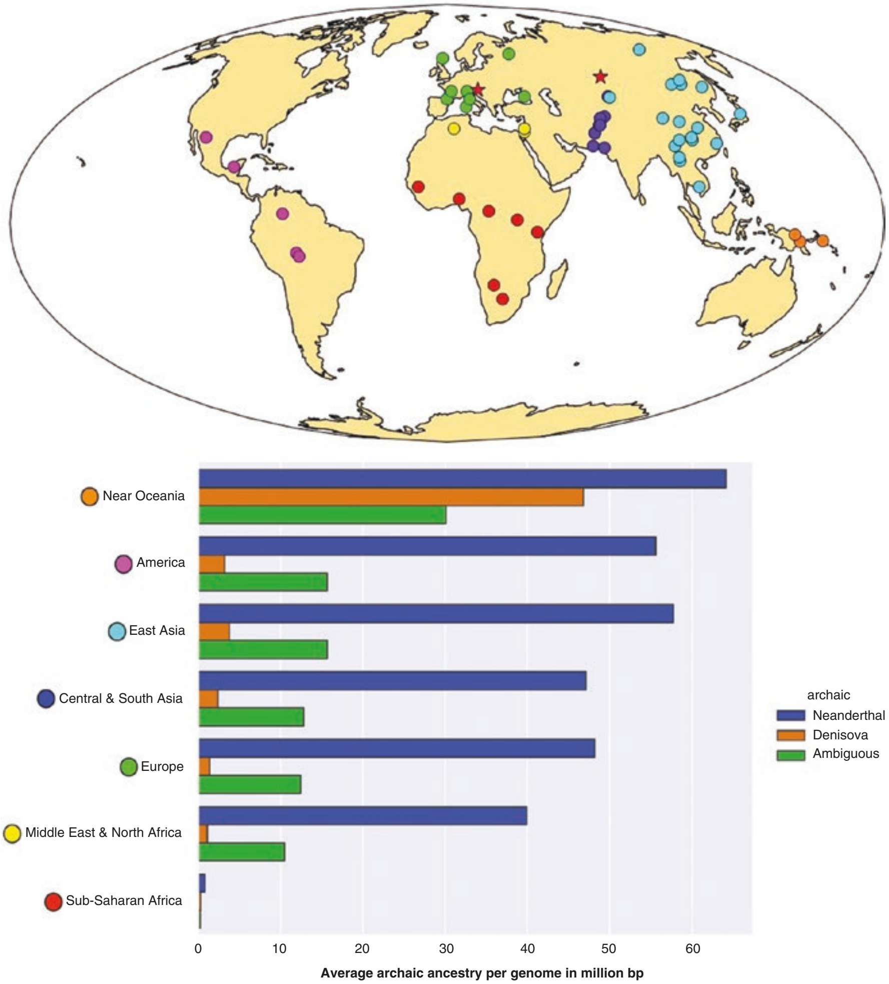


Fig. 11.2 Distribution of variants of archaic hominins in modern human genomes. Continental (non-admixed) populations averages of total haplotype length (in base pairs per diploid genome) with an ancestral source of Neanderthal, Denisovan or either. The 54 different populations analyzed in Bergström et al. [2] were plotted in an equal-area elliptical Mollweide map projection using QGIS, and the two red stars represent the location of the sequenced Neanderthal (Vindija Cave, Croatia) and Denisovan (Denisova Cave, Russia)

An interesting theoretical consequence of introgression is that some Neanderthal and Denisovan variants found in modern humans would be positively selected [53]. Accordingly, when investigating the phenotypic legacy (inferred from genomes) of Neanderthal and Denisovan hominins in modern humans, there is a wide range of phenotypes correlated to archaic ancestry, as lipid metabolism [54], immune response [55–60], reaction to UV light [60, 61], and even high-altitude adaptation in Tibetans [62, 63] and cold weather adaptation in Greenlandic Inuits [64]. Several of those phenotypes are peculiar in which they are both in high-frequency and population specific (Fig. 11.2), hinting at introgression playing an important role in modern human adaptation to environments very different from the sub-Saharan Africa homeland [65]. However, some variants described initially as derived alleles from archaic hominins were recently suggested to be ancestral variants of *Homo sapiens* lost in many populations during the out of Africa bottlenecks [66]. Other phenotypes inferred to be derived from archaic species include also unfavored conditions like higher risk for some diseases as Parkinson's, Chron's, diabetes, lupus, COVID-19, hypercoagulation, celiac disease, neurological disorders, depression and even tobacco usage [61, 65, 67–69]. However, some of these disease-associated variants likely derived from archaic hominins were recently evaluated in an extensive Icelandic pedigree taking flanking variants and African diversity into account [70], which failed to confirm most of the previous disease associations.

In conclusion, while archaic hominin introgression is possibly ubiquitous among all world populations (Fig. 11.2) and taking it into account is necessary to understand modern-day human variation, more methods and ancient genomes are needed to fully comprehend its impact and functional consequences [71]. However, particularly due the higher divergence between archaic and modern human segments, their population specific distribution and high correlation with immune-related genes, archaic derived variants may be key for developing and evaluating the efficacy of some therapeutic strategies in different human groups.

11.4 The Origin and Diversity of Indigenous Populations in Pre-Columbian Times

Since the origin of anatomically modern *Homo sapiens* in Africa at about 200,000 ya, human populations have expanded and settled all remaining continents in the last 70,000 years, giving rise to many indigenous (aborigines, natives) societies distributed worldwide (Fig. 11.1). Indigenous communities can be broadly defined biologically, and genealogically as human populations composed mainly by descendants from ancestors who arrived in the same continental region some few thousand years ago. Thus, all past human populations before the XV century transoceanic navigation enterprises could be called biologically indigenous, whether they were located in Europe, Asia, Africa, Oceania and Americas. Besides, many "relatively" isolated populations (as they were in the XV century) can be still genealogically referred today as indigenous, aborigines, tribal groups or minorities, although other anthropological concepts can be equally—and probably more adequately—applied to define indigenous populations. In the current Cosmopolitan society, indigenous is who is considered and considers itself indigenous [72, 73].

The indigenous societies in every continent were originally resulted from long-term ongoing dispersal events of small hunter-gatherer groups [24, 25], which also retained particular phenotypes under distinctive environmental pressures along many generations submitted to a slight (if not completely isolated) long-distance gene-flow [74]. During their long history of differentiation and adaptation, indigenous groups were also influenced by major cultural changes that impacted the variation at the human genome. For example, since the agricultural beginnings 12,000 ya, human groups started to ingest different types of food that further selected genetic variants throughout many generations of human populations as an adaptive response. For instance, a major impact is found among pastoralist societies who used milk as an important food supply, and substantial changes promoted by natural selection are observed in their genomes [75]. Genomic studies revealed that the *LCT* gene that codes for the lactase enzyme shows the strongest signals of selection in the human genome. In the majority of individuals, *LCT* switches off around the age of 7 years old, which is the ancestral pattern found in all hominids. The variants that allow the persistence of lactase onto adulthood of lactose tolerant individuals are fixed in some populations [74, 76, 77]. Interestingly, while milk was accessible to most populations in Western Eurasia at around 9000 ya, the allele for lactase persistence was still very rare until 4000 ya, reaching high frequency from Portugal to Kazakhstan more recently [27, 78]. While a single *LCT* allele was selected in Europe and Central Asia, other five different *LCT* alleles allow lactase persistence in African and Middle Eastern populations, while displaying a more complex soft selective sweep [79–81].

Because the indigenous period comprised 99.7% of the human history, a large part of the genetic variation of current human societies are derived from this long period of change and adaptation to the environment, food, toxins and pathogens in the history of each particular human population worldwide. For example, among variants accumulated during human prehistory it includes the sickle cell anemia allele (βS) in Africa, the *PDE10A* gene variant found in diving populations of Bajau in Southeast Asia [82], and several variants in genes related to the adaptation to high altitude occurring independently in three highland populations from Tibet, Ethiopia and Andes [83, 84]. The local adaptation of indigenous ancestors also explains why genomic ancestry in modern individuals (in relation to continental/indigenous descent) can be indicative of some health conditions and disease risk [85], and can be also used to map some genes associated to clinically related phenotypes [74, 86].

11.5 Genomic Variation of Contemporaneous Human Societies

The formation of contemporaneous populations, particularly the cosmopolitan urban societies, comprises a complex intermingling of populations that started in the XVI century. Since then, genetic pools have been increasingly admixed and new variants arisen by mutation and recombination accumulated in these last 25 generations or five centuries (Fig. 11.3). Moreover, a hypothesis called the Columbian Exchange [87] raises ideas about an adaptive introgression effect in modern human

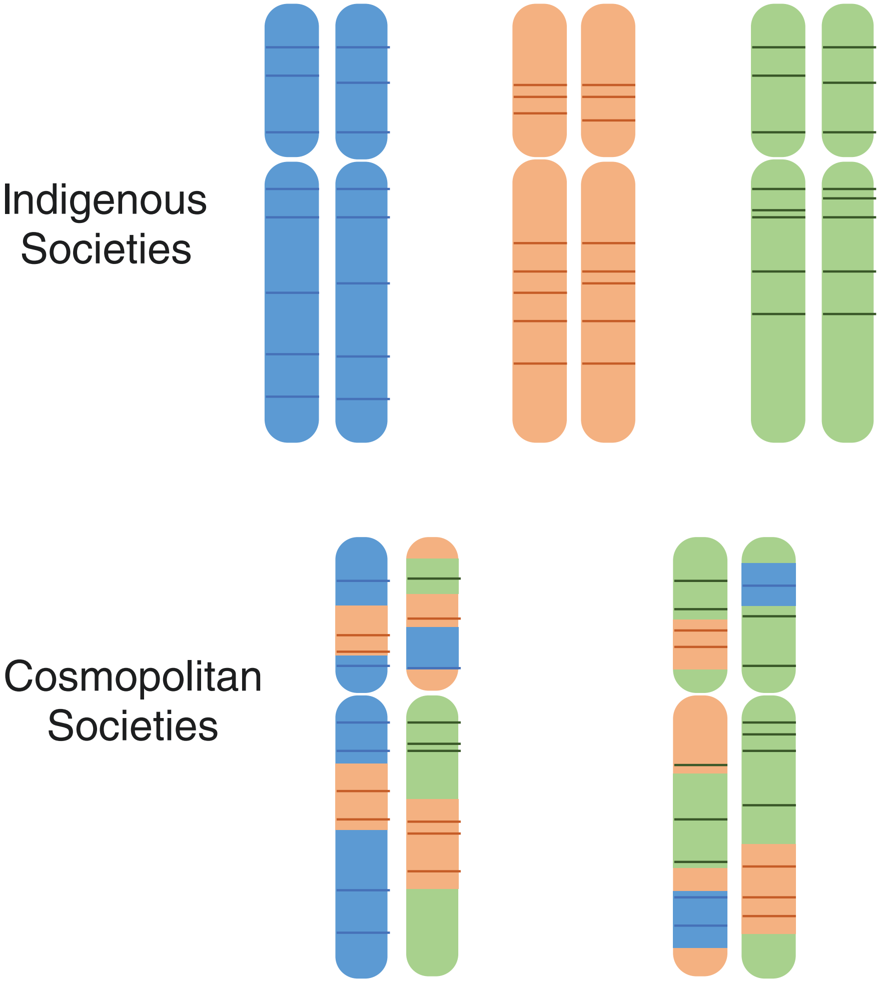


Fig. 11.3 Schematic chromosome architecture in Indigenous and Cosmopolitan societies. Variation (horizontal lines) was heterogeneously distributed between continental indigenous populations until the XV century (different background colors). Modern Cosmopolitan chromosomes are composed by recombinant chromatins with different continental ancestries

populations due to the increasing cultural and biological admixture in the post-Columbian era.

Because of the minor importance of long-distance gene flow in the worldwide populations until the XV century, large blocks of haplotypes with particular variants were usually restricted to continental groups [85]. The intercontinental admixture starting in the post-Columbian age allowed new recombinant chromosomes with an increasing rupture of previously linked variants. Contemporaneous populations are thus composed by individuals with mixed ancestry components, with different combinations of alleles and recombinant segments (Fig. 11.3). Built on this evidence, an admixture-based method has been devised to map diseases or traits using differential risk or susceptibility linked to ancestry [88].

11.6 Variation in Simple and Complex Phenotypes and Inherited Diseases

Huntington's disease was the first clinically relevant phenotype mapped to a human chromosome locus using DNA variation [89]. Since then, millions of DNA variants have been associated to clinical phenotypes by linkage disequilibrium or directly as disease-causing alleles [90].

High throughput genome sequencing techniques are allowing a rapid discovery of causal and susceptibility related variants of many diseases, particularly when one or few candidate genes are involved. However, the majority of common and rare diseases are complex phenotypic traits with multifactorial causes, influenced by genetic (nature) and environment (nurture) variables. Besides, epigenetic variation is also associated with many simple and complex traits, and epistatic effects due to a complex interaction of functional (and variable) molecules are still poorly understood.

Simple, complex, common and rare diseases affected by hereditary traits have been deeply investigated with human population studies using genomic approaches [90, 91], which can be also understood in an evolutionary context [86]. Complex traits associated to human diseases have been successfully mapped using GWAS, which were largely empowered by the use of whole-genome sequencing (WGS) of human populations [92]. Many genes involved in disease etiology are being mapped by the identification of candidate disease-causing or susceptibility-related variants using WGS, GWAS and large-scale bioinformatics analysis [91].

11.7 Variation and Rare Genetic Diseases

Rare disorders affect about 350 million people worldwide and most of them have no formal treatment approved in health agencies like the FDA/USA (https://rarediseases.info.nih.gov). Actually, when treatment is available it is usually too expensive, and drug development is not pursued by pharmaceutical companies for economic reasons. More than 80% of rare disorders has a genetic origin, but no genes can be clearly associated to most of the disease phenotypes. However, rare genetic disorders (or orphan diseases) diagnostics is becoming less expensive with current genomic technologies, which can contribute to prognosis determination of patients and relatives, as well as to indicate an appropriate disease treatment [93].

It has been assumed for many years that rare genetic diseases are usually caused by variants recently originated by mutation events, also called *de novo* mutations. Indeed, it seems to be the case for many rare and common complex neurodevelopmental diseases, including some forms of intellectual disability, autism and schizophrenia [94]. However, many rare disorders are found in different families and sometimes associated to the same genetic variants, which sometimes show a remote origin in the past. Anyway, rare and uncommon genetic disorders can be investigated with help of several databases and bioinformatics tools [93].

11.8 Human Variation in the Prognosis of Pathogenic Diseases

Since our origins in Africa, modern humans have been exposed to very different environmental conditions and a plentiful panoply of exo- and endo pathogens. In some extent, the health and disease of modern populations are consequence of the natural history of different human populations encountering pathogens in newly colonized habitats [86]. As a result of these encounters, selection on human populations exposed to specific pathogens for many generations have favored some alleles associated to pathogenic resistance. For example, some hemoglobin variants, even though causing inherited diseases like sickle cell anemia and thalassemia, are classical paradigms of human genetic traits associated to resistance to Malaria caused by the parasite *Plasmodium falciparum* in Africa and Asia [86]. Many variants associated with susceptibility or resistance to infectious diseases can be traced back to a long history of human contact with specific pathogens [95, 96].

Other pathologies are very recent in human history (less than 100 years), but few variants have also been identified to promote resistance. For example, the Δ32-deletion of the *CCR5* receptor gene precludes the entrance of the HIV virus, which causes AIDS, a zoonotic disease that was "transmitted" (spillover) from African chimpanzees to humans in the XX century [97]. Curiously, this variant has a relatively high prevalence in Europe (5–16.4%), but it is absent in indigenous populations from Africa, Asia and the Americas [98]. Even though Δ32/*CCR5* heterozygous and homozygous individuals have not shown signs of resistance to any other known pathogen, recent studies found a protective action against diabetes type 1, an autoimmune disease injuring the pancreas [99]. Anyway, the example of Δ32/*CCR5* variant promoting individual resistance to AIDS shows how natural selection can change through time (new pathogens), causing alleles that were previously drifting away to become now positively selected (advantageous) in a new environment (HIV infection).

Another viral zoonosis is currently pandemic. The sudden appearance of the SARS-CoV-2 virus causing COVID-19 (Coronavirus Disease 2019) has prompted a global response of scientists to its characterization and to search for its cure or vaccine. This is a new and highly contagious disease in humans that spreads very fast and has arrived in all continents in just few months [100]. The manifestation of

COVID-19 is very variable in the population with a range from asymptomatic individuals to high severity cases leading to death. Severity is associated with age greater than 60 years old, and some comorbidities like obesity, hypertension, diabetes, etc. However, there are rare cases of previously healthy young individuals who presented severe disease and death; in contrast there are elders with comorbidities who remain unaffected despite being heavily exposed to SARS-CoV-2 [100]. This indicates an existence of innate susceptibility or resistance to COVID-19 due to genetic variation, which is being investigated since the beginning of the pandemic [101]. The first candidate genes suspected in human innate response to COVID-19 encode proteins used for SARS-CoV-2 entry into human cells, like *ACE2* and *TMPRSS2*. Indeed, recent studies indicated that *ACE2* and *TMPRSS2* variants may modulate viral infectivity in humans, making some individuals more vulnerable than others [102, 103]. Another host gene named *FURIN* codes for a proprotein convertase that acts together with *TMPRSS2* to cleave the spike protein of the SARS-CoV-2 capsid, enabling attachment and a high affinity association with *ACE2* at the surface of cells. Indeed, specific *FURIN* gene variants have been also linked to an increased risk of contagion by facilitating the entry of the virus in the cells of the respiratory tract [104]. In addition, lower expressions of *ACE2* and *TMPRSS2* genes in African populations have been also associated with protective effects to COVID-19 [102]. Other studies also suggest a relationship between *ACE* deletion allele and reduced *ACE2* expression, thus theoretically it should decrease the probability to be infected by SARS-CoV-2 [105]. Moreover, COVID-19 and other infectious diseases are potentially modulated by genetic variants of the immunity genes (see below), and, for instance, some *in silico* analyses indicated different variants likely associated to protection or vulnerability to SARS-CoV-2 infection [106]. However, patients affected with COVID-19 develop fever, “cytokine storm” and respiratory distress, i.e., a series of complex metabolic responses involving dozens of genes, thus identifying genomic variants associated to predisposal or resistance would be very difficult. More recently, a genome-wide association study analyzing 8,582,968 SNPs in 1980 COVID-19 patients with respiratory failure from Italy and Spain identified variants associated to risk factors in two gene clusters at 3p21.31 and 9q34 [107]. Interestingly, the predisposing variants in the gene cluster at 3p21.31 have been recently suggested to be inherited from Neanderthals [69].

11.9 Human Variation in the Immune Response

The human body presents two defense systems against pathogens that are highly conserved among all vertebrates: innate and adaptive immunity. The innate system is the first line of defense that detects specific molecules (and molecular patterns) associated to pathogens like viruses, bacteria, protozoans, fungus, toxins, etc. [108]. The response of the innate system eliminates pathogens and also prepares the body for the adaptive immunity, the second line of defense. Many genetic variants in the protein genes of the innate systems are potentially associated to predisposition or resistance to infectious diseases. The innate system proteins include *Toll*-like receptors (*TLRs*), nucleotide-binding oligomerization domain (*Nod*), leucine-rich repeat–containing receptors (*NLRs*), *RIG-I* like receptors (*RLRs*), C-type lectin receptors (*CLRs*) and *AIM-2* like receptors, as well as a family of enzymes that function as intracellular sensors of nucleic acids, including *cGAS* and *OAS* proteins [108]. Genetic variants of the innate system have been tested for association with many infectious diseases. For example, the pathogenesis of bacterial meningitis has been linked to variation in many genes involved in innate immunity [109]. Furthermore, two genes related to innate immunity have shown to be consistent with tuberculosis predisposition: *IL12RB1* and *TYK2* [110]. For example, homozygosity of the *TYK2* P1104A variant accounts for 1% of tuberculosis cases in Europeans [111].

The adaptive immune system is mainly composed by proteins encoded by the *MHC* locus producing the glycoproteins *HLA* (Human Leukocyte Antigen) and other proteins [112]. *MHC* loci include around 200 genes grouped mainly in two classes: MHC class I with three main genes *HLA*-A, -B and -C with about two hundred common alleles (>0.001) in these genes; and *MHC* class II with *HLA-DR, -DQ, -DP, -DM, and -DO* with other two hundred common alleles. In addition, MHC class I and class II have shown around 15,000 minor alleles, which in typical heterozygous individual for each of these *MHC* genes, about 1012 different peptides are expressed, increasing the immune defense against any emerging infection. The nature of these proteins is to display the highest panoply of combinations to identify the largest number of antigens [113]. The combination of alleles at *MHC* of any individual is so high that exists a unique set of alleles in every human, except in monozygotic twins. Within the high *MHC* diversity, many genetic variants in particular *MHC* loci have been associated to increased susceptibility or protection to some infectious and autoimmune diseases [112]. For example, aspartic acid at position 57 of the HLA-DQ beta chain is protective [114] against future development of insulin dependent diabetes mellitus (IDDM), and any other aminoacid in this same position, particularly alanine, can predispose to IDDM [115].

The genetic variants of the innate and adaptive systems are still poorly understood, mainly because of the high complexity of genes and variation of the human immune systems. However, new studies suggest that about 20–40% of interindividual variance of the immune response can be explained by genetic variants, known to drive the differential susceptibility to diseases and the vaccine response related particularly to pathogen-caused diseases [116]. Besides, because in much of the human history the indigenous populations experienced different sort of pathogens and environments, part of the variation in immune genes is also expected to differentiate among populations, particularly isolated ones. It seems to be the case of indigenous peoples of Americas, which have a long history of viral epidemics disseminating whole populations since European colonization as discussed above [45]. Indeed, during the H1N1 pandemic of 2009, indigenous peoples in Brazil presented a 4.5-fold higher death rate compared to the country's general population [117]. Furthermore, another study focused in an outbreak of acute respiratory disease in a Guarani indigenous population from southeast Brazil indicated that they display a higher vulnerability to acute respiratory infections and H1N1 vaccination was not effective [118]. A disproportionate impact of influenza was also reported in Australian aborigines [119], Torres Strait islanders [120] and Native Canadians [121]. Even though no specific human variants can be directly associated to vulnerability of these indigenous populations, other studies have characterized the low diversity (and likely low immune response) of indigenous populations from Oceania and Americas as a result of serial founder effects during the initial settlement of the continents out of Africa [122].

11.10 Epigenetic Variation

Human genomic variation is largely connected to epigenetics, influencing gene transcription, chromatin states, genome stability and mutability. Indeed, genotypically identical inherited variation is associated to different disease manifestations because of gender-related epigenetic effects, like Prader-Willi (paternal inheritance) and Angelman (maternal inheritance) syndromes [123]. Besides, many SVs like transposable element insertions are usual targets of epigenetic silencing that can influence gene expression and genome integrity and associated to genetic disorders [124].

The joint investigation of epigenetic components associated to genotypic variation identified by WGS approaches can be used to develop prognostic and diagnostic markers of some human disorders, as well as future targets for therapy [123, 124]. This new discipline, sometimes called epigenomics, is currently using new approaches like CRISPR-based technologies to understand health and disease phenotypes likely caused by epigenetic variants across cell types, tissues and individuals [125].

References

1. Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. High resolution of human evolutionary trees with polymorphic microsatellites. Nature. 1994;368:455–7.
2. Bergström A, McCarthy SA, Hui R, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367:6484.
3. Lander E, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
4. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409:928–33.
5. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851–61.
6. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061.
7. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.

8. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
9. Brookes A, Robinson P. Human genotype–phenotype databases: aims, challenges and opportunities. Nat Rev Genet. 2015;16:702–15. https://doi.org/10.1038/nrg3932.
10. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020;21:171–89. https://doi.org/10.1038/s41576-019-0180-9.
11. Prado-Martinez J, Sudmant PH, Kidd JM, et al. Great ape genetic diversity and population history. Nature. 2013;499:471–5.
12. Stringer C. The origin and evolution of *Homo sapiens*. Philos Trans R Soc B. 2016;371:20150237. https://doi.org/10.1098/rstb.2015.0237.
13. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, Munson KM, Hastie AR, Diekhans M, Hormozdiari F, Lorusso N, Hoekzema K, Qiu R, Clark K, Raja A, Welch AE, Sorensen M, Baker C, Fulton RS, Armstrong J, Graves-Lindsay TA, Denli AM, Hoppe ER, Hsieh P, Hill CM, Pang AWC, Lee J, Lam ET, Dutcher SK, Gage FH, Warren WC, Shendure J, Haussler D, Schneider VA, Cao H, Ventura M, Wilson RK, Paten B, Pollen A, Eichler EE. High-resolution comparative analysis of great ape genomes. Science. 2018;360:eaar6343. https://doi.org/10.1126/science.aar6343.
14. Azevedo L, Serrano C, Amorim A, et al. Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. Hum Genomics. 2015;9:21. https://doi.org/10.1186/s40246-015-0043-1.
15. Grabowski M, Jungers WL. Evidence of a chimpanzee-sized ancestor of humans but a gibbon-sized ancestor of apes. Nat Commun. 2017;8:880. https://doi.org/10.1038/s41467-017-00997-4.
16. Fisher SE. Human genetics: the evolving story of FOXP2. Curr Biol. 2019;29:R65–7. https://doi.org/10.1016/j.cub.2018.11.047.
17. Montgomery SH, Capellini I, Venditti C, Barton RA, Mundy NI. Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. Mol Biol Evol. 2011;28:625–38. https://doi.org/10.1093/molbev/msq237.
18. Kuhlwilm M, Boeckx C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. Sci Rep. 2019;9:8463. https://doi.org/10.1038/s41598-019-44877-x.
19. Pinotti T, Bergström A, Geppert M, Bawn M, Ohasi D, Shi W, Lacerda DR, Solli A, Norstedt J, Reed K, Dawtry K, González-Andrade F, Paz-Y-Miño C, Revollo S, Cuellar C, Jota MS, Santos JE, Ayub Q, Kivisild T, Sandoval JR, Fujita R, Xue Y, Roewer L, Santos FR, Tyler-Smith C. Y chromosome sequences reveal a short Beringian standstill, rapid expansion, and early population structure of native American founders. Curr Biol. 2019;29:149–57.
20. Tylén K, Fusaroli R, Rojo S, Heimann K, Fay N, Johannsen NN, Riede F, Lombard M. The evolution of early symbolic behavior in *Homo sapiens*. Proc Natl Acad Sci U S A. 2020;117:4578–84. https://doi.org/10.1073/pnas.1910880117.
21. McHugo GP, Dover MJ, MacHugh DE. Unlocking the origins and biology of domestic animals using ancient DNA and paleogenomics. BMC Biol. 2019;17(1):98. https://doi.org/10.1186/s12915-019-0724-7.
22. Piperno DR. Assessing elements of an extended evolutionary synthesis for plant domestication and agricultural origin research. Proc Natl Acad Sci U S A. 2017;114:6429–37. https://doi.org/10.1073/pnas.1703658114.
23. Diamond J, Bellwood P. Farmer and their languages: the first expansions. Science. 2003;300:597–603.
24. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343:747–51. https://doi.org/10.1126/science.1243518.
25. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017;541:302–10.
26. Willerslev E, Cooper A. Ancient DNA. Proc Royal Soc B: Biol Sci. 2005;272:3–16.

27. Allentoft M, Sikora M, Sjögren K-G, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522:167–72.
28. Haak W, Lazaridis I, Patterson N, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522:207–11.
29. Lazaridis I, Patterson N, Mittnik A, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513:409–13.
30. Olalde I, Mallick S, Patterson N, et al. The genomic history of the Iberian Peninsula over the past 8000 years. Science. 2019;363:1230–4.
31. McColl H, Racimo F, Vinner L, et al. The prehistoric peopling of Southeast Asia. Science. 2018;361:88–92.
32. Damgaard PB, Martiniano R, Kamm J, et al. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science. 2018;360:6396.
33. Lazaridis I, Nadel D, Rollefson G, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536:419–24.
34. Sikora M, Pitulko VV, Sousa VC, et al. The population history of northeastern Siberia since the Pleistocene. Nature. 2019;570:182–8.
35. Haber M, Mezzavilla M, Xue Y, Tyler-Smith C. Ancient DNA and the rewriting of human history: be sparing with Occam's razor. Genome Biol. 2016;17:1–8.
36. Prohaska A, Racimo F, Schork AJ, Sikora M, Stern AJ, Ilardo M, Allentoft ME, Folkersen L, Buil A, Moreno-Mayar JV, Korneliussen T, Geschwind D, Ingason A, Werge T, Nielsen R, Willerslev E. Human disease variation in the light of population genomics. Cell. 2019;177:115–31.
37. Hellwege JN, Keaton JM, Giri A, Gao X, Edwards DR, Edwards TL. Population stratification in genetic association studies. Curr Protoc Hum Genet. 2017;95:1.22.1–1.22.23. https://doi.org/10.1002/cphg.48.
38. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999;65:220–8.
39. Alves-Silva J, Santos MSS, Guimarães PEM, Ferreira ACS, Bandelt H-J, Pena SDJ, Prado VF. The ancestry of Brazilian mtDNA lineages. Am J Hum Genet. 2000;67:444–61.
40. Carvajal-Carmona LG, Soto ID, Pineda N, Ortíz-Barrientos D, Duque C, Ospina-Duque J, McCarthy M, Montoya P, Alvarez VM, Bedoya G, Ruiz-Linares A. Strong Amerind/White Sex Bias and a possible Sephardic contribution among the founders of a population in Northwest Colombia. Am J Hum Genet. 2000;67:1287–95.
41. Carvalho-Silva DR, Santos FR, Rocha J, Pena SDJ. The phylogeography of Brazilian Y-chromosome lineages. Am J Hum Genet. 2001;68:281–6.
42. Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martínez-Fuentes A, Comas D. Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol. 2008;8:213.
43. Moreno-Mayar JV, Vinner V, Damgaard PB, et al. Early human dispersals within the Americas. Science. 2018;362:6419.
44. Posth C, Nakatsuka N, Lazaridis I, et al. Reconstructing the deep population history of Central and South America. Cell. 2018;175:1185–97.
45. Denevan WM. The pristine myth: the landscape of the Americas 1492. Ann Am Geographers. 1992;82:369–85.
46. Green RE, et al. A draft sequence of the Neandertal genome. Science. 2010;328:710–22.
47. Prüfer K, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;35:655–8.
48. de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P, Schmidt JM, Heredia-Genestar JM, Benazzo A, Barbujani G, Peter BM, Kuderna LFK, Casals F, Angedakin S, Arandjelovic M, Boesch C, Kühl H, Vigilant L, Langergraber K, Novembre J, Gut M, Gut I, Navarro A, Carlsen F, Andrés AM, Siegismund HR, Scally A, Excoffier L, Tyler-Smith C, Castellano S, Xue Y, Hvilsom C,

Marques-Bonet T. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354:477–81.

49. Meyer M, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;328:222–6.
50. Reich D, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468:1053–60.
51. Durvasula A, Sankararaman S. Recovering signals of ghost archaic introgression in African populations. Sci Adv. 2020;6(7):eaax5097. https://doi.org/10.1126/sciadv.aax5097.
52. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci U S A. 2011;108:15123–8.
53. Harris K, Nielsen R. The genetic cost of Neanderthal introgression. Genetics. 2016;203:881–91.
54. Khrameeva EE, Bozek Z, He L, Yan Z, Jiang X, Wei Y, Tang K, Gelfand MS, Prüfer K, Kelso J, Pääbo S, Giavalisco P, Lachmann M, Khaitovich P. Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. Nat Commun. 2014;5:1–8.
55. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, Kimani J, Carrington M, Middleton D, Rajalingam R, Beksac M, Marsh SG, Maiers M, Guethlein LA, Tavoularis S, Little AM, Green RE, Norman PJ, Parham P. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science. 2011;334:89–94.
56. Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, Chen Y, Hurles ME, Tyler-Smith C, Xue Y. Population structure, stratification, and introgression of human structural variation. Cell. 2020;182:189–99.
57. Dannemann M, Andrés AM, Kelso J. Introgression of Neandertal- and Denisovan-like haplotypes contributes to adaptive variation in human Toll-like receptors. Am J Hum Genet. 2016;98:22–33.
58. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova J-L, Patin E, Quintana-Murci L. Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am J Hum Genet. 2016;98:5–21.
59. Quach H, Rotival M, Pothlichet J, Loh Y-HE, Dannemann M, Zidane N, Laval G, Patin E, Harmant C, Lopez M, et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell. 2016;167:643–56.
60. Vernot B, Akey JM. Resurrecting surviving Neandertal lineages from modern human genomes. Science. 2014;343:1017–21.
61. Dannemann M, Kelso J. The contribution of Neanderthals to phenotypic variation in modern humans. Am J Hum Genet. 2017;101:578–89. https://doi.org/10.1016/j.ajhg.2017.09.010.
62. Arciero E, Kraaijenbrink T, Asan HM, Mezzavilla M, Ayub Q, Wang W, Pingcuo Z, Yang H, Wang J, Jobling MA, van Driem G, Xue Y, de Knijff P, Tyler-Smith C. Demographic history and genetic adaptation in the Himalayan region inferred from genome-wide SNP genotypes of 49 populations. Mol Biol Evol. 2018;35:1916–33.
63. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, Ni P, Wang B, Ou X, Huasang, Luosang J, Cuo ZXP, Li K, Gao G, Yin Y, Wang W, Zhang X, Xu X, Yang H, Li Y, Wang L, Wang J, Nielsen R. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512:194–7.
64. Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, Albrechtsen A, Carmel L, Huerta-Sánchez E, Nielsen R. Archaic adaptive introgression in TBX15/WARS2. Mol Biol Evol. 2017;34:509–24.
65. Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. Archaic hominin admixture facilitated adaptation to out-of-Africa environments. Curr Biol. 2016;26:3375–82. https://doi.org/10.1016/j.cub.2016.10.041.
66. Rinker DC, Simonti CN, McArthur E, Shaw D, Hodges E, Capra JA. Neanderthal introgression reintroduced functional ancestral alleles lost in Eurasian populations. Nat Ecol Evol. 2020; https://doi.org/10.1038/s41559-020-1261-z.

67. Sankararaman S, Mallick S, Patterson N, Reich D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Curr Biol. 2016;26:1241–7. https://doi.org/10.1016/j.cub.2016.03.037.
68. Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, Crosslin DR, Hebbring SJ, Jarvik GP, Kullo IJ, Li R, Pathak J, Ritchie MD, Roden DM, Verma SS, Tromp G, Prato JD, Bush WS, Akey JM, Denny JC, Capra JA. The phenotypic legacy of admixture between modern humans and Neandertals. Science. 2016;351:737–41.
69. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature. 2020;587(7835):610–2. https://doi.org/10.1038/s41586-020-2818-3.
70. Skov L, Maciá MC, Svelnbjörnsson G, Mafessoni F, Lucotte EA, Einarsdóttir MS, Jonsson H, Haldorsson B, Gudbjartsson DF, Helgason A, Schierup MH, Stefansson K. The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. Nature. 2020;582:78–83.
71. Dannemann M, Racimo F. Something old, something borrowed: admixture and adaptation in human evolution. Curr Opin Genet Dev. 2018;53:1–8.
72. Barth F. Ethnic groups and boundaries. The social organization of culture difference. Oslo: Universitetforlaget; 1969.
73. Carneiro da Cunha M. Etnicidade: da cultura residual mas irredutível. Revista de Cultura e Política. 1986;1:35–9.
74. Fan S, Hansen ME, Lo Y, Tishkoff SA. Going global by adapting local: a review of recent human adaptation. Science. 2016;354:54–9. https://doi.org/10.1126/science.aaf5098.
75. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, Pritchard JK. Detection of human adaptation during the past 2000 years. Science. 2016;354:760–4.
76. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004;74:1111–20.
77. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. Nat Genet. 2002;30:233–7.
78. Gamba C, Jones ER, Teasdale MD, et al. Genome flux and stasis in a five millennium transect of European prehistory. Nat Commun. 2014;5:5257.
79. Jones BL, Raga TO, Liebert A, Zmarz P, Bekele E, Danielsen ET, Olsen AK, Bradman N, Troelsen JT, Swallow DM. Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. Am J Hum Genet. 2013;93:538–44.
80. Macholdt E, Lede V, Barbieri C, Mpoloka SW, Chen H, Slatkin M, Pakendorf B, Stoneking M. Tracing pastoralist migrations to Southern Africa with lactase persistence alleles. Curr Biol. 2014;24:875–9.
81. Tishkoff SA, Reed FA, Ranciaro A, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007;39:31–40.
82. Ilardo MA, Moltke I, Korneliussen TS, et al. Physiological and genetic adaptations to diving in Sea Nomads. Cell. 2018;173:569–580.e15. https://doi.org/10.1016/j.cell.2018.03.054.
83. Beall CM. Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. Integr Compar Biol. 2006;46:18–24. https://doi.org/10.1093/icb/icj004.
84. Witt KE, Huerta-Sánchez E. Convergent evolution in human and domesticate adaptation to high-altitude environments. Philos Trans R Soc B. 2019;374:20180235. https://doi.org/10.1098/rstb.2018.0235.
85. Norris ET, Wang L, Conley AB, Rishishwar L, Mariño-Ramírez L, Valderrama-Aguirre A, Jordan IK. Genetic ancestry, admixture and health determinants in Latin America. BMC Genomics. 2018;19:861. https://doi.org/10.1186/s12864-018-5195-7.
86. Quintana-Murci L. Understanding rare and common diseases in the context of human evolution. Genome Biol. 2016;7:225.
87. Jordan IK. The Columbian exchange as a source of adaptive introgression in human populations. Biol Direct. 2016;11:17. https://doi.org/10.1186/s13062-016-0121-x.

88. Skotte L, Jørsboe E, Korneliussen TS, Moltke I, Albrechtsen A. Ancestry-specific association mapping in admixed populations. Genet Epidemiol. 2019;43:506–21. https://doi.org/10.1002/gepi.22200.
89. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983;306:234–8.
90. Hitomi Y, Tokunaga K. Significance of functional disease-causal/susceptible variants identified by whole-genome analyses for the understanding of human diseases. Proc Jpn Acad Ser B Phys Biol Sci. 2017;93:657–76. https://doi.org/10.2183/pjab.93.042.
91. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111:6131–8. https://doi.org/10.1073/pnas.1318948111.
92. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet. 2015;16:275–84. https://doi.org/10.1038/nrg3908.
93. Pogue RE, Cavalcanti DP, Shanker S, Andrade RV, Aguiar LR, de Carvalho JL, Costa FF. Rare genetic diseases: update on diagnosis, treatment and online resources. Drug Discov Today. 2017;23:187–95. https://doi.org/10.1016/j.drudis.2017.11.002.
94. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012;13:565–75. https://doi.org/10.1038/nrg3241.
95. Chapman S, Hill A. Human genetic susceptibility to infectious disease. Nat Rev Genet. 2012;13:175–88. https://doi.org/10.1038/nrg3114.
96. Klebanov N. Genetic Predisposition to Infectious Disease. Cureus. 2018;10:e3210. https://doi.org/10.7759/cureus.3210.
97. Faria NR, Rambaut A, Suchard MA, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science. 2014;346:56–61. https://doi.org/10.1126/science.1256739.
98. Solloch UV, Lang K, Lange V, Böhme I, Schmidt AH, Sauter J. Frequencies of gene variant CCR5-Δ32 in 87 countries based on next-generation sequencing of 1.3 million individuals sampled from 3 national DKMS donor centers. Hum Immunol. 2017;78:710–7.
99. Słomiński B, Tawrynowicz U, Ryba-Stanisławowska M, Skrzypkowska M, Myśliwska J, Myśliwiec M. CCR5-Δ32 polymorphism is a genetic risk factor associated with dyslipidemia in patients with type 1 diabetes. Cytokine. 2019;114:81–5. https://doi.org/10.1016/j.cyto.2018.11.005.
100. Pascarella G, Strumia A, Piliego C, et al. COVID-19 diagnosis and management: a comprehensive review. J Intern Med. 2020;288:192–206. https://doi.org/10.1111/joim.13091.
101. The COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur J Hum Genet. 2020;28:715–8. https://doi.org/10.1038/s41431-020-0636-6.
102. Ortiz-Fernández L, Sawalha AH. Genetic variability in the expression of the SARS-CoV-2 host cell entry factors across populations. Genes Immunol. 2020;21:269–72. https://doi.org/10.1038/s41435-020-0107-7.
103. Torre-Fuentes L, Matías-Guiu J, Hernández-Lorenzo L, et al. ACE2, TMPRSS2, and Furin variants and SARS-CoV-2 infection in Madrid. Spain J Med Virol. 2020;93(2):863–9. https://doi.org/10.1002/jmv.26319.
104. Shang J, Wan Y, Luo C, Ye G, Geng Q, Li F. Cell entry mechanisms of SARS-CoV-2. PNAS. 2020;117:11727–34. https://doi.org/10.1073/pnas.2003138117.
105. Delanghe JR, Speeckaert MM, De Buyzere ML. ACE polymorphism and COVID-19 outcome. Endocrine. 2020;70:13–4. https://doi.org/10.1007/s12020-020-02454-7.

106. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, Thompson RF. Human leukocyte antigen susceptibility map for severe acute respiratory syndrome coronavirus 2. J Virol. 2020;94:e00510–20. https://doi.org/10.1128/JVI.00510-20.
107. The Severe COVID-19 GWAS Group. Genomewide association study of severe covid-19 with respiratory failure. N Engl J Med. 2020; https://doi.org/10.1056/NEJMoa2020283.
108. Iwasaki A, Medzhitov R. Control of adaptive immunity by the innate immune system. Nat Immunol. 2015;16:343–53.
109. Sanders M, van Well G, Ouburg S, et al. Genetic variation of innate immune response genes in invasive pneumococcal and meningococcal disease applied to the pathogenesis of meningitis. Genes Immun. 2011;12:321–34. https://doi.org/10.1038/gene.2011.20.
110. Boisson-Dupuis S. The monogenic basis of human tuberculosis. Hum Genet. 2020;139:1001–9.
111. Kerner G, Ramirez-Alejo N, Seeleuthner Y, Yang R, Ogishi M, Cobat A, Patin E, Quintana-Murci L, Boisson-Dupuis S, Casanova JL, Abel L. Homozygosity for TYK2 P1104A underlies tuberculosis in about 1% of patients in a cohort of European ancestry. Proc Natl Acad Sci U S A. 2019;116:10430–4. https://doi.org/10.1073/pnas.1903561116.
112. Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. Genome Biol. 2017;18:76. https://doi.org/10.1186/s13059-017-1207-1.
113. Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, Moraes ME, Pereira SE, Kempenich JH, Reed EF, Setterholm M, Smith AG, Tilanus MG, Torres M, Varney MD, Voorter CE, Fischer GF, Fleischhauer K, Goodridge D, Klitz W, Little AM, Maiers M, Marsh SG, Müller CR, Noreen H, Rozemuller EH, Sanchez-Mazas A, Senitzer D, Trachtenberg E, Fernandez-Vina M. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. Tissue Antigens. 2013;81:194–203. https://doi.org/10.1111/tan.12093.
114. Boehm BO, Manfras B, Rosak C, Schöffling K, Trucco M. Aspartic acid at position 57 of the HLA-DQ beta chain is protective against future development of insulin-dependent (type 1) diabetes mellitus. Klin Wochenschr. 1991;69:146–50. https://doi.org/10.1007/BF01665854.
115. Reinauer C, Rosenbauer J, Bächle C, et al. The clinical course of patients with preschool manifestation of type 1 diabetes is independent of the HLA DR-DQ genotype [published correction appears in Genes (Basel)]. Genes. 2018;8(5):146. https://doi.org/10.3390/genes8050146.
116. Liston A, Carr EJ, Linterman MA. Shaping variation in the human immune system. Trends Immunol. 2016;37:637–46. https://doi.org/10.1016/j.it.2016.08.002.
117. La Ruche G, Tarantola A, Barboza P, Vaillant L, Gueguen J, Gastellu-Etchegorry M, for the epidemic intelligence team at InVS. The 2009 pandemic H1N1 influenza and indigenous populations of the Americas and the Pacific. Euro Surveill. 2009;14:19366.
118. Cardoso AM, Resende PC, Paixao ES, et al. Investigation of an outbreak of acute respiratory disease in an indigenous village in Brazil: Contribution of Influenza A(H1N1)pdm09 and human respiratory syncytial viruses. PLoS One. 2019;14:e0218925. https://doi.org/10.1371/journal.pone.0218925.
119. Flint SM, Davis JS, Su JY, Oliver-Landry EP, Rogers BA, Goldstein A, et al. Disproportionate impact of pandemic (H1N1) 2009 influenza on Indigenous people in the Top End of Australia's Northern Territory. Med J Aust. 2010;192:617–22.
120. Trauer JM, Laurie KL, McDonnell J, Kelso A, Markey PG. Differential effects of pandemic (H1N1) 2009 on remote and indigenous groups, Northern Territory, Australia, 2009. Emerg Infect Dis. 2011;17:1615–23.
121. Pollock SL, Sagan M, Oakley L, Fontaine J, Poffenroth L. Investigation of a pandemic H1N1 influenza outbreak in a remote First Nations community in northern Manitoba, 2009. Can J Public Health. 2012;103:90–3.
122. Henn BM, Cavalli-Sforza LL, Feldman MW. The great human expansion. Proc Natl Acad Sci U S A. 2012;109:17758–64. https://doi.org/10.1073/pnas.1212380109.

123. Zink F, Magnusdottir DN, Magnusson OT, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. Nat Genet. 2018;50:1542–52. https://doi.org/10.1038/s41588-018-0232-7.

124. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. Nature. 2019;571:489–99. https://doi.org/10.1038/s41586-019-1411-0.

125. Stricker SH, Köferle A, Beck S. From profiles to function in epigenomics. Nat Rev Genet. 2017;18:51–66. https://doi.org/10.1038/nrg.2016.138.

Index

L. A. Haddad (ed.), *Human Genome Structure, Function and Clinical Considerations*, https://doi.org/10.1007/978-3-030-73151-9