Kyriacos Felekkis
Konstantinos Voskarides *Editors*

Genomic Elements in Health, Disease and Evolution

Junk DNA

Springer

Genomic Elements in Health, Disease and Evolution

Kyriacos Felekkis
Konstantinos Voskarides
Editors

Genomic Elements in Health, Disease and Evolution

Junk DNA

Springer

Editors

Kyriacos Felekkis
Department of Life and Health Sciences and
University of Nicosia Medical School
University of Nicosia
Nicosia, Cyprus

Konstantinos Voskarides
Department of Biological Sciences
Molecular Medicine Research Center
University of Cyprus
Nicosia, Cyprus

ISBN 978-1-4939-3069-2 ISBN 978-1-4939-3070-8 (eBook)
DOI 10.1007/978-1-4939-3070-8

Library of Congress Control Number: 2015950479

Springer New York Heidelberg Dordrecht London

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Preface

Although the term "junk" DNA was used since the early 1960s, the term's origin was attributed to Susumo Ohno who officially used the term to describe pseudogenes' sequences resulted from gene duplication and subsequent mutagenesis events. Since then, the term was widely used to describe any non-coding sequence of the genome. Today, "junk" DNA refers to any genomic sequence that does not play a functional role in the organism. The use of the term was accompanied by various unanswered questions: Why do we have so much "junk" DNA in our genome? Do these non-coding sequences have functional significance? The discovery of novel genomic elements in the recent years was a step forward in an attempt to address these issues. It appears that the percentage of the non-functional DNA is being significantly reduced as more and more functions are attributed to those non-coding regions of the genome. Despite the continuous shrinkage of the non-functional portion of the genome, it is believed that a significant part of the genome is indeed non-functional.

In this book, we attempt to provide a thorough review of various non-coding genomic elements and discuss in depth their role in health, disease and evolution. We begin our exploration with non-coding RNA molecules, miRNAs, piRNAs, LncRNAs and transposable elements as these moieties dominate the scientific literature in the last 10 years. We proceed with the discussion of copy number variation regions, mini- and micro-satellites, and proximal and distal elements of the genome. The last section of this book focuses on the review of well-known non-coding regions of the genome, introns, centromeres and telomeres, but enriched with newly discovered functions. As the vast amount of data in regard to these elements is attributed to a great degree to the growing technology in the field of biomedicine, the last chapter of this book discusses the latest development in the field of Next Generation Sequence and the potential applications of this technology in the study of non-coding regions of the genome.

The original structure of this book was greatly shaped by many conversations with colleagues in Cyprus and abroad. We are indebted to all the authors contributing to this publication for their in-depth review of the subject and their excellent writing. We must also thank all the scientists whose work is included in this book.

Special thanks to our respective institutions and colleagues for their support and critical interventions. The driving force of our inspiration is equivocally our students who provide us with the strength and willingness to sit down and write. Last but not least, we are indebted to our families and friends for their continuous support and encouragement.

Nicosia, Cyprus

Kyriacos Felekkis
Konstantinos Voskarides

Contents

1 **MiRNAs' Function and Role in Evolution: Under the View of Genomic Enhancement Phenomena** 1
Konstantinos Voskarides and Kyriacos Felekkis

2 **MicroRNAs in Disease** 17
Gregory Papagregoriou

3 **piRNAs-Transposon Silencing and Germ Line Development** 47
Catherine Demoliou

4 **Long Non-coding RNA** 83
Monika Gullerova

5 **Unveiling Transposable Elements Function to Enrich Knowledge for Human Physiology and Disease Pathogenesis** 109
Ioannis S. Vizirianakis, Elsa P. Amanatiadou, and Sotirios S. Tezias

6 **Copy Number Variation in Human Health, Disease and Evolution** 129
Carolina Sismani, Costas Koufaris, and Konstantinos Voskarides

7 **Mini- and Micro-Satellite Markers in Health, Disease and Evolution** 155
Vasiliki A. Galani, Sofia Markoula, Leandros Lazaros,
Paris Ladias, and Ioannis Georgiou

8 **Intron Biology, Focusing on Group II Introns, the Ancestors of Spliceosomal Introns** 195
María Dolores Molina-Sánchez, Rafael Nisa-Martínez,
Fernando M. García-Rodriguez, Francisco Martínez-Abarca,
and Nicolás Toro

9 **Centromeres in Health, Disease and Evolution** 221
Thian Thian Beh and Paul Kalitsis

10 Structure and Functions of Telomeres in Organismal Homeostasis and Disease 247
Penelope Kroustallaki and Sarantis Gagos

11 Proximal Regulatory Elements with Emphasis on CpG Rich Regions 285
Pavlos Fanis

12 Genomic Analysis Through High-Throughput Sequencing 297
Michalis Hadjithomas

Index 313

Contributors

Elsa P. Amanatiadou, M.Sc. Laboratory of Pharmacology, Department of Pharmaceutical Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

Thian Thian Beh, B.Sc. (Hons.) Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, VIC, Australia

Department of Paediatrics, University of Melbourne, Melbourne, VIC, Australia

Catherine Demoliou, Ph.D. Life and Health Sciences Department, School of Sciences and Engineering, University of Nicosia, Nicosia, Cyprus

Pavlos Fanis, B.Sc., Ph.D. Molecular Genetics Thalassaemia Department, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus

Kyriacos Felekkis, B.Sc., Ph.D. Department of Life and Health Sciences, University of Nicosia Medical School, University of Nicosia, Nicosia, Cyprus

Sarantis Gagos, Ph.D. Department of Experimental Medicine and Translational Research, Biomedical Research Foundation of the Academy of Athens (BRFAA), Athens, Greece

Vasiliki A. Galani, Ph.D. Department of Anatomy – Histology – Embryology, University of Ioannina, Ioannina, Epirus, Greece

Fernando M. García-Rodriguez, Ph.D. Grupo de Ecología Genética, Microbiología del Suelo y Sistemas Simbióticos, Estación Experimental del Zaidín/ Consejo Superior de Investigaciones Científicas, Granada, Spain

Ioannis Georgiou, Ph.D. Division of Reproductive Genetics, Department of Obstetrics & Gynecology, University of Ioannina, Ioannina, Epirus, Greece

Division of Medical Genetics and Clinical Embryology, Department of Obstetrics & Gynecology, University Hospital of Ioannina, Ioannina, Epirus, Greece

Monika Gullerova, Ph.D. Sir William Dunn School of Pathology, University of Oxford, Oxford, UK

Michalis Hadjithomas, B.Sc., Ph.D. DOE Joint Genome Institute, Walnut Creek, CA, USA

Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Paul Kalitsis, Ph.D. Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, VIC, Australia

Department of Paediatrics, University of Melbourne, Melbourne, VIC, Australia

Costas Koufaris, Ph.D. Department of Cytogenetics and Genomics, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus

Penelope Kroustallaki, B.Sc., M.Sc. Department of Experimental Medicine and Translational Research, Biomedical Research Foundation of the Academy of Athens (BRFAA), Athens, Greece

Paris Ladias, B.Sc. Division of Reproductive Genetics, Department of Obstetrics & Gynecology, University of Ioannina, Ioannina, Epirus, Greece

Leandros Lazaros, Ph.D. Division of Reproductive Genetics, Department of Obstetrics & Gynecology, University of Ioannina, Ioannina, Greece

Laboratory of Medical Genetics and Clinical Embryology, Department of Obstetrics & Gynecology, University Hospital of Ioannina, Ioannina, Epirus, Greece

Sofia Markoula, M.D., Ph.D. Department of Neurology, University Hospital of Ioannina, Ioannina, Epirus, Greece

Francisco Martínez-Abarca, Ph.D. Grupo de Ecología Genética, Microbiología del Suelo y Sistemas Simbióticos, Estación Experimental del Zaidín/Consejo Superior de Investigaciones Científicas, Granada, Spain

María Dolores Molina-Sánchez, Ph.D. Grupo de Ecología Genética, Microbiología del Suelo y Sistemas Simbióticos, Estación Experimental del Zaidín/Consejo Superior de Investigaciones Científicas, Granada, Spain

Rafael Nisa-Martínez, Ph.D. Grupo de Ecología Genética, Microbiología del Suelo y Sistemas Simbióticos, Estación Experimental del Zaidín/Consejo Superior de Investigaciones Científicas, Granada, Spain

Gregory Papagregoriou, Ph.D. Department of Biological Sciences, Molecular Medicine Research Center, University of Cyprus, Nicosia, Cyprus

Carolina Sismani, Ph.D. Department of Cytogenetics and Genomics, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus

Sotirios S. Tezias, M.Sc. Laboratory of Pharmacology, Department of Pharmaceutical Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

Nicolás Toro, Ph.D. Grupo de Ecología Genética, Microbiología del Suelo y Sistemas Simbióticos, Estación Experimental del Zaidín/Consejo Superior de Investigaciones Científicas, Granada, Spain

Ioannis S. Vizirianakis, Ph.D. Laboratory of Pharmacology, Department of Pharmaceutical Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

Konstantinos Voskarides, Ph.D. Department of Biological Sciences, Molecular Medicine Research Center, University of Cyprus, Nicosia, Cyprus

Chapter 1
MiRNAs' Function and Role in Evolution: Under the View of Genomic Enhancement Phenomena

Konstantinos Voskarides and Kyriacos Felekkis

Introduction: Biogenesis, Nomenclature and the Principal Role of miRNAs

MiRNAs are short, single-stranded RNA molecules approximately 21–23 nucleotides in length (mature miRNAs), which usually have a uridine at their 5′ end and they are partially complementary to one or more messenger RNA (mRNA) molecules. Although they cover only a small part of the genome, their role in gene expression regulation is considered to be of high significance. MiRNAs were discovered by Victor Ambros, Rosalind Lee and Rhonda Feinbaum in 1993 during a study of the gene *lin-14* in *C. elegans* development [1]. They found that LIN-14 protein abundance was regulated by a short RNA product encoded by the *lin-4* gene, which was partially complementary to many regions of the *lin-14* 3′ UTR. This complementarity was sufficient to inhibit the translation of the *lin-14* mRNA and necessary for the worm development. A huge amount of data has been accumulated since then for miRNAs and many more are yet to be discovered in the near future.

Due to this great and continuous flow of data, a specific nomenclature system has been created. The prefix "mir" is followed by a dash and a number, the latter often indicating order of naming. For example, mir-123 was named and likely discovered prior to mir-456. The uncapitalized "mir-" refers to the pre-miRNA, while a capitalized "miR-" refers to the mature form. MiRNAs with nearly identical sequences bar one or

K. Voskarides, Ph.D. (✉)
Department of Biological Sciences, Molecular Medicine Research Center,
University of Cyprus, Kallipoleos 75, 1678 Nicosia, Cyprus
e-mail: kvoskar@ucy.ac.cy

K. Felekkis, B.Sc., Ph.D. (✉)
Department of Life and Health Sciences, University of Nicosia Medical School,
University of Nicosia, 46 Makedonitissas Ave., 1700 Nicosia, Cyprus
e-mail: Felekkis.k@unic.ac.cy

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_1

two nucleotides are annotated with an additional lower case letter. For example, miR-123a would be closely related to miR-123b. Pre-miRNAs that lead to 100 % identical mature miRNAs but that are located at different locations in the genome are indicated with an additional dash-number suffix. For example, the pre-miRNAs hsa-mir-194-1 and hsa-mir-194-2 lead to an identical mature miRNA (hsa-miR-194) but are located in different regions of the genome. Species of origin is usually designated with a three-letter prefix, e.g., hsa-miR-123 is a human (*Homo sapiens*) miRNA and oar-miR-123 is a sheep (*Ovis aries*) miRNA. When relative expression levels are known, an asterisk following the name indicates a miRNA that is expressed at low levels relative to the miRNA in the opposite arm of a hairpin. For example, miR-123 and miR-123* would share a pre-miRNA hairpin, but more miR-123 would be found in the cell.

But what is known about the "molecular play" of miRNAs? Each miRNA is thought to regulate multiple genes. Since hundreds of miRNA genes are predicted to be present in higher eukaryotes, the potential regulatory circuitry afforded by miRNAs is vast [2, 3]. Acting at the post-transcriptional level, these molecules may regulate the expression of more than 30 % of all mammalian protein-coding genes [4]. Several research groups have verified that miRNAs may act as key regulators of processes as diverse as embryonic development, cell proliferation, cell growth, tissue differentiation and apoptosis. The mature miRNA can bind even with partial complementarity to the target mRNA (typically on 3′ UTR), downregulating the translation of the mRNA. The mature miRNA mainly acts by targeting a miRNA recognition element (MRE) on the mRNA's 3′ UTR and binds on it through a Watson–Crick base-pairing manner. MiRNA target recognition properties depend on its 'seed region'. Recognition binding sequences are short, usually 6–8 nt [5–8]. The expression of a large number of the predicted human miRNA genes has been confirmed, but many predicted miRNA targets remain to be identified and verified (reviewed in Chap. 2).

Most of the miRNA genes are located in "spacing" DNA, the DNA that is found between different genes. A percentage of 40 % of miRNA genes may lie in the introns of protein and non-protein coding genes or even in exons of long non protein-coding transcripts [9]. These are usually, though not exclusively, found in a sense orientation [10, 11] and are usually regulated together with their host genes [9, 12, 13]. A number of miRNA genes have a common promoter, including 42–48 % of all polycistronic miRNAs (found in the same genetic region), containing multiple discrete loops from which mature miRNAs are processed [14, 15]. This does not necessarily mean the mature miRNAs of a family will have similar structure and functions. The promoters mentioned have some similarities in their motifs to promoters of other genes transcribed by RNA polymerase II, such as protein coding genes [14, 16]. In brief, animal miRNAs are transcribed by RNA polymerase II and processed by the microprocessor protein complex (DGCR8/Drosha) into precursor stem-loop miRNAs in the nucleus. The cleavage of a pri-miRNA by microprocessor begins with DGCR8 recognizing the ssRNA–dsRNA junction typical of a pri-miRNA. Drosha is approaching its substrate through interaction with DGCR8 and cleaves the stem of a pri-miRNA ~11 nt away from the two single stranded segments (~22 nts away from the loop) (summarized by Felekkis et al. [17]). Drosha removes the double-stranded stem from the remainder of the pri-miRNA by cleaving proximal and distal of the stem, generating a pre-miRNA that has a 5′-monophosphate and a 3′-2-nt overhang (Fig. 1.1).

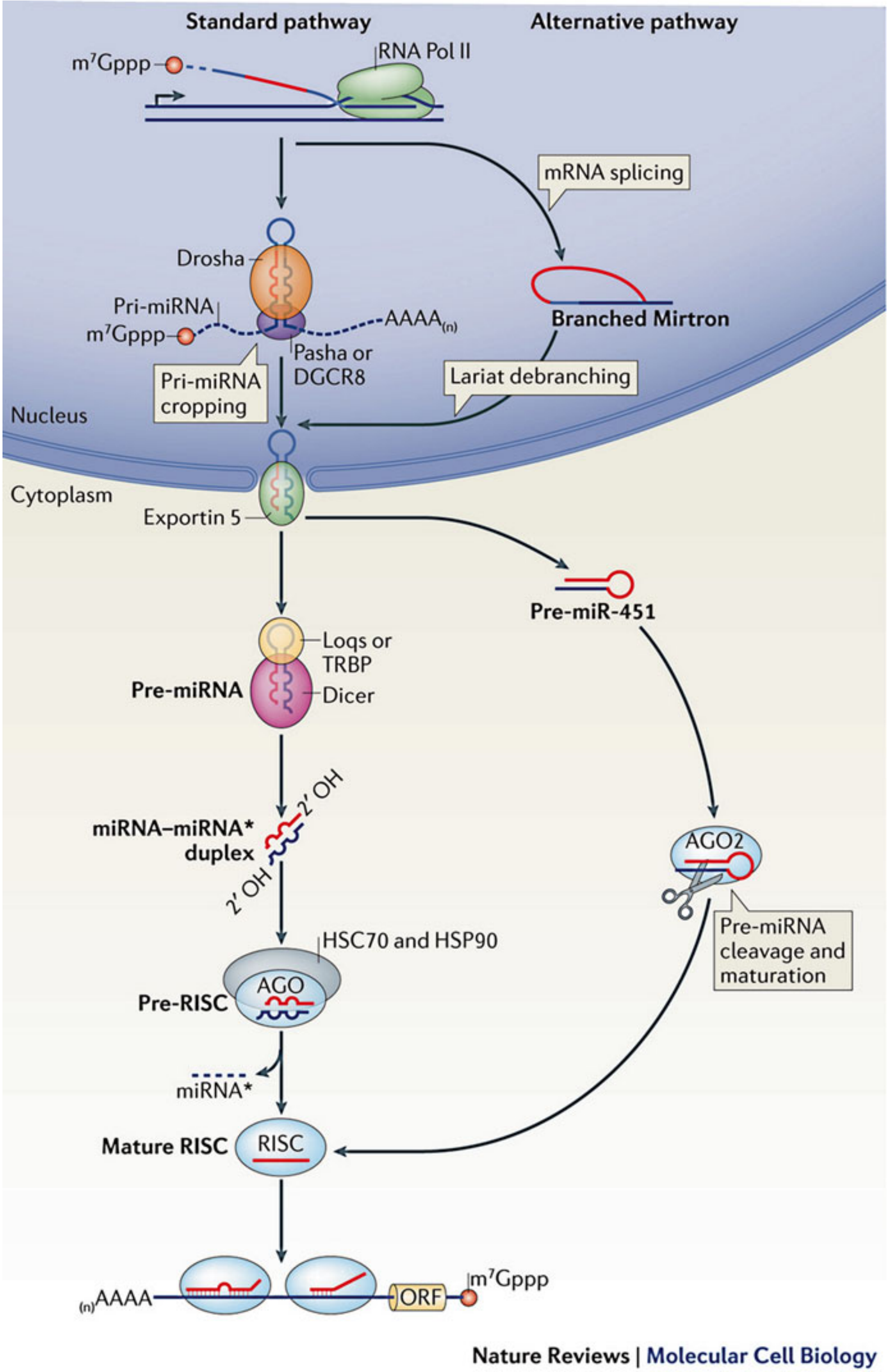


Fig. 1.1 Several layers of regulation control miRNA gene biogenesis in animals: transcription activation, splicing, recognition by Drosha, post-processing, RNA editing, sub-cellular localization, nuclear export, and hairpin arm selection. From Ameres SL, Zamore PD. Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol. 2013 Aug; 14(8):475–488. Reprinted with permission from Nature Publishing Group

Precursor miRNAs (pre-miRNAs) are exported into the cytoplasm by Exportin-5 and are cleaved by Dicer to produce mature miRNAs. After cleavage, the miRNA duplex is unwound by an unidentified RNA helicase and the mature miRNA strand binds to an Argonaute (Ago) protein into an RNPcomplex (RSIC) (Fig. 1.1). The other miRNA strand is degraded. A primary determinant as to which of the two strands of a miRNA duplex will be loaded on Ago proteins is the inherent thermodynamic asymmetry of the miRNA duplex. The RNA strand whose 5′- end is less stably bound to the opposite strand will be loaded to Ago proteins and form the mature miRNA [17].

Mature miRNAs recognize their respective target mRNAs and mediate post-transcriptional repression of their targets through translational repression, deadenylation, or enhanced mRNA decay. The procedure is similar in plants, with the difference that plants have no Drosha homolog. The plant Dicer homolog, DICER-LIKE1 (DCL1), orchestrates both processing events within the nucleus, typically resulting in an 21-nucleotide mature miRNA/miRNA passenger strand duplex with two-nucleotide 39 overhangs [18–20].

MiRNA Genes Evolution

Chromosomal Organization of miRNA Genes

An interesting observation is that miRNA gene number per chromosome correlates with the protein-coding gene density. This indicates that integration and/or maintenance of miRNA genes roughly follows protein-coding genes. However, *Homo Sapiens* chromosomes 14, 19, and X are exceptionally enriched for miRNA genes, something that may be related with the evolutionary history of those chromosomes.

As previously mentioned miRNAs that are found in genetic clusters are called polycistronic and tend to have a common promoter. It is obvious that those miRNAs were formed though repeated genomic duplication events caused by the unequal recombination—the not perfect alignment of homologous chromosomes during meiosis—a very frequent phenomenon in chromosomal regions with genetic repeats. The end result of such event will be the gain or loss of a single copy.

Plausibly, employment of an already existing functional promoter by new miRNA genes is an efficient way to express new miRNAs, eliminating the need for de novo establishment of promoter-enhancer sequences upstream of the miRNA gene. The subsequent result of this phenomenon is miRNAs (in regions up to ~50 kb) that tend to be co-expressed [12, 15]. Amplification of an ancestral miRNA inside a cluster is not only significant for “de novo” birth of novel miRNAs but possibly contributes to the effective dosage of a given expressed miRNA homolog [11, 15].

Old and Young miRNA Genes

Conserved miRNA sequences among species, are considered as the oldest ones. Newly derived miRNAs are the non-conserved miRNAs and are derived through duplication events and subsequent nucleotide substitutions [21]. Frequency of duplication of derived miRNAs may be different in plants and animals since there is evidence that hairpin structure is the main source of new miRNAs in *Drosophila* species instead of duplication events that dominate in miRNAs' birth in plants [22, 23]. Nozawa et al. [22] found evidence that miRNAs are initially controlled by neutral evolution but once miRNA genes acquired their functions, they appear to evolve very slowly, maintaining essentially the same structures for a long time. This suggests that once a miRNA gains a function it undergoes extreme purifying selection. On the other hand, miRNAs which are more recently derived (and thus presumably non-functional) are frequently lost (Fig. 1.2).

A lot of interest has been displayed on functional differences among old and young miRNAs. Comparative genomics suggests that ancient miRNAs have on average two fold more targets than newly generated ones [24]. On the other hand, new miRNAs genes were shown to have lower expression compared with the old ones [25]. Expression regulation of intronic miRNAs (intragenic) is of considerable interest. Do they follow the expression pattern of their host gene? A recent publication provides evidence that expression of intragenic miRNAs differs between young and old miRNAs. Young intragenic miRNAs display lower levels of co-expression with host genes than old ones [26]. This interesting co-evolutionary relation requires further investigation.

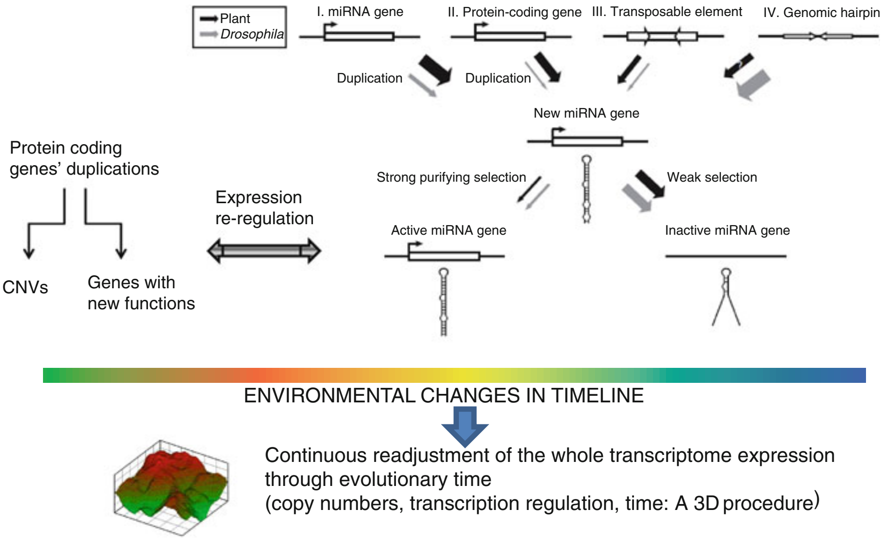

Plant
Drosophila
I. miRNA gene
II. Protein-coding gene
III. Transposable element
IV. Genomic hairpin
Duplication
Duplication
New miRNA gene
Protein coding genes' duplications
CNVs
Genes with new functions
Expression re-regulation
Strong purifying selection
Weak selection
Active miRNA gene
Inactive miRNA gene
ENVIRONMENTAL CHANGES IN TIMELINE
Continuous readjustment of the whole transcriptome expression through evolutionary time
(copy numbers, transcription regulation, time: A 3D procedure)

Fig. 1.2 Model of miRNAs as regulators of duplication phenomena in evolution. From Nozawa M, Miura S, Nei M (2012) Origins and evolution of miRNA genes in plant species. Genome Biol Evol 4:230–239. With permission from Oxford University Press

Without a doubt, the main question is "when did the first miRNAs appeared and why?" A way to answer this question is to identify the old (ancient) miRNAs and study their functions. Comparative sequence data indicate that the oldest known animal miRNA is miR-100. Related to this miRNA the miR-125 and let-7 that were initially active in neurosecretory cells located around the mouth. Other sets of ancient miRNAs were first present in locomotor ciliated cells, specific brain centers, or, more broadly, one of four major organ systems: central nervous system, sensory tissue, musculature and gut. These findings reveal that miRNA evolution and the establishment of tissue identities were closely coupled in bilaterian evolution [27].

But what about plants? Combined with numerous data sets from high-throughput sequencing experiments, eight miRNA families have been identified in the common ancestor of all embryophytes. The MIR396 family was present in the common ancestor of all tracheophytes (vascular plants), while the MIR397 and MIR398 families were acquired in the common ancestor of all spermatophytes. A high proportion of species-specific or non-conserved miRNA genes were also observed in various plant species. This demonstrates that many plant miRNAs have been raised during speciation processes (for more information see the detailed review of Cuperus et al. [28]).

Genomic Duplications, Repetitive Regions and Transposons (TEs) as Potential Sources of miRNA Genes

A duplication can be segmental (from a few nucleotides to several thousand kilobases—"tandem" is a similar term used for exactly the same and connected genetic repeats) or may cover the whole genome (an event also called polyploidization). Segmental duplication (or small-scale duplication) and polyploidization correspond to distinct evolutionary processes with widely different impacts. Another type of frequent duplications in genomes is the inverted duplications. These are segmental and inversely repeated genetic regions.

Just as with protein-encoding genes, a major route for new miRNA genes evolution is genomic duplications. Segmental or inverted duplications can be a source of new miRNA genes. There is substantial evidence that many of the miRNA families originated through multiple genomic duplication events or through repetitive genetic elements [2, 21, 29–33]. Several molecular mechanisms can determine the future of such new miRNA genes. Sub-functions or completely new functions can be "born". We encourage the readers who are interested in all the evolutionary processes regarding miRNAs to read the detailed review "The evolution and functional diversification of animal miRNA genes" by Liu et al. [34].

Tandem duplications can result in paralogous miRNA sequences that are located on the same transcript and organized as tandem paralog clusters [31, 35]. In a recent study, evidence is given that repetitive elements contribute to the de novo origin of miRNAs in mammalian genomes and that large segmental duplication events accelerate the expansion of miRNA families, including those derived from repetitive sequences [33]. The latter ones are considered as the younger miRNA genes, being also the less conserved among species. Sun et al. [36] found similar evidence in plants, underlining that differences exist between the miRNAs that are found in repetitive elements and those that are not. The former tend to have longer hairpin precursor, lower G-C content in hairpin precursor sequences and lower minimum free energy.

Based on the observation that some miRNA precursors had extended similarity beyond miRNA sequences with target genes, Allen et al. [37] proposed the inverted duplication model. Under this hypothesis, new miRNA genes are generated from inverted duplication events happened on one of their target genes by forming two adjacent gene segments in either convergent or divergent orientation. Recent observations showing that a large proportion of miRNA genes are included in transposable elements (TEs) or pseudogenes, urged scientists to believe that inverted duplications are closely related with TEs (see next paragraph) or pseudogenes. Zhang et al. [38] further confirmed that the inverted duplication model in plants is happening via TEs or pseudogenes, showing also that inverted duplications give rise to miRNAs much more frequently that segmental duplications.

Data regarding the evolution of miRNAs from TEs are significant and continuously accumulating. Smalheiser and Torvik [39] reported 11 instances of presumably TE-derived mammalian miRNAs. Later studies by Piriyapongsa and Jordan [40] and Piriyapongsa et al. [41] identified that 12 % of miRNAs in their study data set overlap with TEs in the human genome. These miRNAs reside within TE copies of all four major TE classes including short interspersed repetitive elements (SINEs), long interspersed repetitive elements (LINEs), long terminal repeats (LTRs), and DNA transposons, suggesting that the formation of novel miRNAs from these elements has occurred multiple times during the human genome evolution. Devor [42] demonstrated seven marsupial-specific miRNAs that possibly evolved from marsupial-specific TEs. Recently, strong evidence was published for similar procedures in plants [32, 36, 43]. The very detailed bioinformatic analysis by Li et al. [32] gave results supporting the notion that TEs in gene rich regions in plants can form foldbacks in non-coding part of transcripts that may eventually evolve into miRNA genes or be integrated into protein coding sequences to form potential targets in a "temperate" manner. A similar work by Sun et al. [36] in four plant species confirmed that a significant number of miRNAs in plants derived from TEs. In addition, Lenhert et al. [44] found evidence in mice that lineage-specific retrotransposons have played an important role in the birth of new miRNA genes during evolution.

Dahary et al. [45] published more evidence on TEs and miRNAs evolution, giving also data for the first time in regards to the relationship of CpG islands (CG repeats) with miRNAs. They found that 300 bp upstream and downstream of the miRNA gene, the observed-to-expected ratio of CpG is significantly higher than the genome average. Further analysis identified 65 human miRNAs that overlap CpG islands (59 of these are fully contained within CpGs) and that none of them were TE associated. The authors believe that the association between miRNAs and CpG islands raises two possible scenarios. Either CpG-rich regions serve as genomic material for miRNAs to emerge or that miRNAs are preserving these regions from their natural decay by methylation and deamination.

MiRNAs Target Sites' Evolution

Co-evolution of miRNA Genes and Their Targets

As mentioned before, a single miRNA appears to control many protein-coding genes. But are the miRNA's target seed sites conserved among different species? Hertel et al. [31] recently reported that human, *Drosophila melanogaster*, and *Caenorhabditis elegans* share 20 common miRNA genes in their genomes. However, bioinformatic analysis suggested that only five miRNA genes share the same target genes among these species [46]. Additionally, as mentioned above, ancient miRNAs have on average twofold more targets than newly generated ones [24]. This finding indicates that most of the miRNAs have experienced gains and losses of their target genes during evolution.

Miura et al. [47] investigated the target sites of miRNAs in Hox genes (the genes that control the body segmentation of metazoan embryos), in 12 *Drosophila* species and in *Daphnia*. Phylogenetic analysis of target sites in Abd-A, Ubx, and Antp Hox genes showed that the old target sites, which existed before the divergence of the 12 Drosophila species, have been well maintained in most species under purifying selection. By contrast, new target sites, which were generated during Drosophila evolution, were often lost in some species and mostly located in non-conserved regions of the 3′ UTRs. These results indicate that these regions can be a potential source for new target sites and in this way creating targets in multiple genes for each miRNA in animals.

More complex factors may contribute to the dual evolution of miRNAs and their targets. The recent and very interesting study of Chen et al. [48] showed that the number of miRNA types that regulate a gene is the strongest indicator of proteins' and genes' evolution. They also divided proteins into low and high intrinsically disordered proteins, finding this way differences in their evolution rates. Additionally, they found that phosphorylated proteins tend to have a higher level of miRNA regulation in their genes and that the number of phosphorylation sites of a protein is correlated with the level of miRNA regulation in low intrinsically disordered proteins. Many different scenarios can explain these results, demonstrating the evolutionary complexity of miRNAs genesis.

Duplications as a Major Evolutionary Pressure for miRNAs' Target Sites Emergence and Their Genomic Distribution

A crucial matter in miRNAs evolution is which factors determine the type and the number of miRNAs' targets in 3′ UTRs. Very recently, it was demonstrated that genomic duplications events constitute the main evolutionary factor for this process. Firstly, Ha et al. [49] proved that small RNAs produced during interspecific mating or polyploidization serve as a buffer against the genomic shock in interspecific hybrids and allopolyploids. The authors came to this conclusion after studying allotetraploids coming from *A. thaliana* and *A. arenosa*, identifying adoptive alterations of the miRNAs and siRNAs levels in comparison with the parental species. Abrouk et al. [50] found evidence that the above mechanism may be a standard procedure in plants after euploidy, especially in euploidy events that are involved in evolutionary speciation. They further suggested that miRNAs may be implicated in genes participating in stress responses pathways, which are essential for plant adaptation and useful for crop variety innovation.

On the other hand, whole genome duplication events are rare in animals and in other phylogenetic clades. Despite this, genetic dosage effects of smaller scale are present in such species. Importantly, Lehnert et al. [51] showed that sense Alu repeat sequences are enriched for miRNA target sites. Even more noteworthy, Li et al. [52] and D'Antonio and Ciccrelli [53] found that miRNA targets are significantly enriched for paralogs genes. Characteristically, Li et al. [52] mention that their results suggest that "miRNA-mediated regulation plays an important role in the regulatory circuits involving duplicated genes including adjusting imbalanced dosage effects of gene duplicates, and possibly creating a mechanism for genetic buffering". On the other hand, D'Antonio and Ciccrelli [53] found that this fine tuning is more significant for "ohnologs", i.e. the paralogs genes that came through vertebrate-specific whole genome duplication events. A more complicate analysis by Fernandez and Chen [54] revealed that human paralogs of poorly packed proteins (categorized so according to special structural criteria) are more likely to be targeted by miRNAs, thus underscoring a means to buffer dosage imbalance effects arising from gene duplication.

Our team provided further evidence for this "genomic duplication" hypothesis. By performing *in silico* whole genome analysis, we demonstrate that both the number of miRNAs that target genes found in Copy Number Variations (CNVs) regions as well as the number of miRNA-binding sites are significantly higher than those of genes found in non-CNV regions [55]. In addition, by examining the miRNA–CNV genes interactions in eight different species we demonstrated that there appears to be an evolutionary dependence on gene expression regulation by miRNAs [56]. There is significant indication that a number of genes located within CNVs have increased (or sometimes reduced) expression level [57–60]. This suggests that miRNAs may have acted as equilibrators of gene expression during evolution in an attempt to regulate aberrant gene expression and to increase the tolerance to genome plasticity. Our results were further confirmed by Woodwark and Bateman [61]. These data raise the possibility that miRNAs may have been created under evolutionary pressure, as a mechanism for increasing the tolerance to genome plasticity.

One specific example that shows the role of miRNAs in tuning the gene dosage of paralogs is represented by atrophins, a phylogenetically conserved family of transcriptional regulators that appeared in metazoans (Atro) and duplicated in vertebrates (ATN1 and Rere). The dosage of the fly atrophin gene Atro is under the tight control of miR-8 [62]. The lack of miR-8 produces Atro overexpression and results in elevated apoptosis in the brain, behavioral defects and severe defects in animal survival [62, 63]. Additionally, reduced Atro expression causes impaired survival, indicating that the fine-tuning dosage of this gene is crucial for its activity [62].

A rarer phenomenon is the targeting of miRNAs on coding sequences. Recent data suggest that in such cases, target sites arise from highly repeated sequences inside the ORFs. The same researchers showed that such sequence repeats largely arise through duplications and occur particularly frequently within families of paralogous C2H2 zinc-finger genes [64].

Adjustment of Expression Levels' Variability According Current Environmental Frame

Wu et al. [65] summarize a number of studies to support the idea that miRNAs are vital molecules for natural selection that proceeds through canalizing evolution. Canalization refers to the process by which phenotypes are macro-evolutionary stabilized within species. The existence and identity of canalizing genes have thus been an important, but controversial topic. For example, Rutherford and Lindquist [66] showed that hsp90 gene may be one of those genes, since its deletion in *Drosophila* can result in many abnormalities, in contrast to "cryptic" genetic variations that are found on it and do not cause any observable phenotype. Hsp90 is an important chaperone, regulating the folding of many proteins. In the same way, miRNAs affect the expression of many genes of the genome.

Despite the evidence for the importance of gene regulation by miRNAs, the typical magnitude of observed repression by miRNAs is relatively small [65, 67, 68]. Wu et al. [65] cite a number of previous studies to show that miRNAs can be functionally classified in two categories. The first category sets the mean of the expression level of the target genes (referred to as expression tuning) and the second one reduces their variance (expression buffering, or homeostasis). On the other hand, some miRNAs' deletions seem not to be "significant" at the organism level. Additionally, decay of some miRNAs genes is considered to be accomplished fast, maybe within millions of years only. Of course, a controversial observation is the conservation over species of a number of miRNA genes and their targets. All these advocate to "canalizing" phenotypes, regulated by miRNAs.

Despite the attractiveness of Wu's ideas [65], we believe that the real situation for miRNAs' evolution is slightly different. "Canalizing" evolution is of course a very important evolutionary phenomenon, since this way significant phenotypic frameworks are secured over the danger of drastic evolutionary changes (like new mutations' emergence, founder effects, genetic drift, occasional directional or adaptive evolution etc.). But in case of miRNAs, we have a kind of a very "flexible" and "adjustable" canalizing evolution. A whole network of gene regulation has been invented by nature in a way that genome can adapt very easily in current environmental changes; but of course without any drastic changes of the main phenotypes. Adding to this equation the phenomenon of genomic duplications, we have then an even larger network, increasing even more the genomic plasticity and "evolvability". In this way, maybe the majority of population members can adopt effectively under an environment change. Alternatively, conventional evolution knowledge demands the elimination of a big part of the population by natural selection, resulting to the survival of few of the best adopted individuals. Such catastrophic evolutionary procedures are highly energy consuming and can slow down evolution for just some relatively mild environmental changes. Sometimes, the only thing needed is just a fine tuning re-regulation of the expression network and this can be achieved effectively by the miRNAs'.

Going back again to genomic duplications, we can suggest that that genomic duplications are the "fuels" and miRNAs are the "regulators" of the genomic engine. An interesting question is what came first, duplications or miRNAs? Gu et al. [69] found evidence that support the evolutionary scenario that gene/genome duplications in the early stage of vertebrates expand the protein-encoding genes and miRNAs simultaneously. The fact is that this is a very complex co-regulation. Adding to this the parameter of evolutionary time, then we can actually have a three dimensional evolutionary process (Fig. 1.2).

Additionally, we must have in mind that environment is not only the one outside the organism boundaries but it is also the one outside the cell boundaries (usually termed as "internal environment"). Under this framework we can examine the results of Mukherji et al. [70] where they found that repression through miRNAs varies a lot among different individual cells. Additionally, they found that a miRNA can behave both as a switch, in the target expression regime below the threshold, and as a fine-tuner, in the sensitive transition between the threshold and the minimal repression regime at high mRNA levels. So, do miRNAs finally regulate their function at the cellular level and not at the organismic level? We previously commented on duplication events as significant factors for miRNAs' evolution. Interestingly, it is now well established that somatically derived CNVs may be an important factor of genetic differentiation among same type of cells. In this concept, evolution in cellular level micro-environments can be the reason of different proteomic profiles that are found in similar cell types and can explain the results of Mukherji et al. [70] and those of other studies. This hypothesis is more plausible in rapidly proliferating cells, like bone marrow stem cells and cancer cells. Selection can act more efficiently through a big number of cell cycles and very possibly new miRNA genes or target sites can emerge or decay this way. Future studies are needed and especially "wet lab" experiments to test these theoretical assumptions.

Conclusions

Nobody can doubt that miRNAs are significant buffering tools that can modulate problems associated with drastic changes in gene expression. It would not be an exaggeration to suggest that miRNAs could have emerged under the pressure of genomic duplications, in order to control "genomic boundaries" and "genomic expression". It is also possible that miRNAs emerged multiple times through evolution (there are some data supporting this hypothesis in animal and plants), for the above reason. What is certain is that since they emerged, they are continuously gaining significant cellular, developmental and evolutionary roles. On the other hand, it would be a unreasonable to believe that this is the whole story in gene expression regulation. We believe that we still have a lot to learn regarding gene regulation and the associated mechanisms.

References

1. Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854
2. Berezicov E, Guryev V, van de Belt J, Wienholds E, Plasterk HAR, Cuppen E (2005) Phylogenetic shadowing and computational identification of human miRNA genes. Cell 120:21–24
3. Xie X, Lu J, Kulbokas EJ, Golub RT, Mootha V, Lindblad-Toh K, Lander SE, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434:338–345
4. Lewis B, Burge C, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are miRNA targets. Cell 120:15–20
5. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian miRNA targets. Cell 115:787–798
6. Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of miRNA-target recognition. PLoS Biol 3:e85
7. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N (2005) Combinatorial miRNA target predictions. Nat Genet 37:495–500
8. Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB (2007) Determinants of targeting by endogenous and exogenous miRNAs and siRNAs. RNA 13:1894–1910
9. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of mammalian miRNA host genes and transcription units. Genome Res 14:1902–1910
10. Cai X, Hagedorn CH, Cullen BR (2004) Human miRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10:1957–1966
11. Weber MJ (2005) New human and mouse miRNA genes found by homology search. FEBS J 272:59–73
12. Baskerville S, Bartel DP (2005) Microarray profiling of miRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA 11:241–247
13. Kim YK, Kim VN (2007) Processing of intronic miRNAs. EMBO J 26:775–783
14. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN (2004) MiRNA genes are transcribed by RNA polymerase II. EMBO J 23:4051–4060
15. Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H (2005) Clustering and conservation patterns of human miRNAs. Nucleic Acids Res 33:2697–2706
16. Zhou X, Ruan J, Wang G, Zhang W (2007) Characterization and identification of miRNA core promoters in four model species. PLoS Comput Biol 3:e37
17. Felekkis K, Touvana E, Stefanou C, Deltas C (2010) miRNAs: a newly described class of encoded molecules that play a role in health and disease. Hippokratia 14:236–240
18. Park W, Li J, Song R, Messing J, Chen X (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in miRNA metabolism in Arabidopsis thaliana. Curr Biol 12:1484–1495
19. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MiRNAs in plants. Genes Dev 16:1616–1626

20. Xie Z, Kasschau KD, Carrington JC (2003) Negative feedback regulation of Dicer-Like1 in Arabidopsis by miRNA-guided mRNA degradation. Curr Biol 13:784–789
21. Tanzer A, Stadler PF (2004) Molecular evolution of a miRNA cluster. J Mol Biol 339:327–335
22. Nozawa M, Miura S, Nei M (2010) Origins and evolution of miRNA genes in Drosophila species. Genome Biol Evol 2:180–189
23. Nozawa M, Miura S, Nei M (2012) Origins and evolution of miRNA genes in plant species. Genome Biol Evol 4:230–239
24. Shomron N, Golan D, Hornstein E (2009) An evolutionary perspective of animal miRNAs and their targets. J Biomed Biotechnol 2009:594738
25. Lu C et al (2008) Genome-wide analysis for discovery of rice miRNAs reveals natural antisense miRNAs (nat-miRNAs). Proc Natl Acad Sci U S A 105:4951–4956
26. He C, Li Z, Chen P, Huang H, Hurst LD, Chen J (2012) Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene coevolution. Nucleic Acids Res 40:4002–4012
27. Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, Klaus S, Snyman H, Hannon GJ, Bork P, Arendt D (2010) Ancient animal miRNAs and the evolution of tissue identity. Nature 25:1084–1088
28. Cuperus JT, Fahlgren N, Carrington JC (2011) Evolution and functional diversification of MIRNA genes. Plant Cell 23:431–442
29. Maher C, Stein L, Ware D (2006) Evolution of Arabidopsis miRNA families through duplication events. Genome Res 16:510–519
30. Jiang D, Yin C, Yu A, Zhou X, Liang W et al (2006) Duplication and expression analysis of multicopy miRNA gene family members in Arabidopsis and rice. Cell Res 16:507–518
31. Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A et al (2006) The expansion of the metazoan miRNA repertoire. BMC Genomics 7:25
32. Li Y, Li C, Xia J, Jin Y (2011) Domestication of transposable elements into MiRNA genes in plants. PLoS One 6:e19212
33. Yuan Z, Sun X, Jiang D, Ding Y, Lu Z et al (2010) Origin and evolution of a placental-specific miRNA family in the human genome. BMC Evol Biol 10:346
34. Liu N, Okamura K, Tyler DM, Phillips MD, Chung WJ, Lai EC (2008) The evolution and functional diversification of animal miRNA genes. Cell Res 18:985–996
35. Zhang L, Chia JM, Kumari S, Stein JC, Liu Z et al (2009) A genome-wide characterization of miRNA genes in maize. PLoS Genet 5:e1000716
36. Sun J, Zhou M, Mao Z, Li C (2012) Characterization and evolution of miRNA genes derived from repetitive elements and duplication events in plants. PLoS One 7:e34092
37. Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW et al (2004) Evolution of miRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet 36:1282–1290
38. Zhang Y, Jiang WK, Gao LZ (2011) Evolution of miRNA genes in Oryza sativa and Arabidopsis thaliana: an update of the inverted duplication model. PLoS One 6:e28073
39. Smalheiser NR, Torvik VI (2005) Mammalian miRNAs derived from genomic repeats. Trends Genet 21:322–326
40. Piriyapongsa J, Jordan IK (2007) A family of human miRNA genes from miniature inverted-repeat transposable elements. PLoS One 2:e203
41. Piriyapongsa J, Marino-Ramirez L, Jordan IK (2007) Origin and evolution of human miRNAs from transposable elements. Genetics 176:1323–1337
42. Devor EJ (2006) Primate miRNAs miR-220 and miR-492 lie within processed pseudogenes. J Hered 97:186–190
43. Piriyapongsa J, Jordan IK (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. RNA 14:814–821
44. Lehnert S, Kapitonov V, Thilakarathne PJ, Schuit FC (2011) Modeling the asymmetric evolution of a mouse and rat-specific miRNA gene cluster intron 10 of the Sfmbt2 gene. BMC Genomics 23:257

45. Dahary D, Shalgi R, Pilpel Y (2011) CpG Islands as a putative source for animal miRNAs: evolutionary and functional implications. Mol Biol Evol 28:1545–1551
46. Chen K, Rajewsky N (2006) Deep conservation of miRNA-target relationships and 3′ UTR motifs in vertebrates, flies, and nematodes. Cold Spring Harb Symp Quant Biol 71:149–156
47. Miura S, Nozawa M, Nei M (2011) Evolutionary changes of the target sites of two miRNAs encoded in the Hox gene cluster of Drosophila and other insect species. Genome Biol Evol 3:129–139
48. Chen SC, Chuang TJ, Li WH (2011) The relationships among miRNA regulation, intrinsically disordered regions, and other indicators of protein evolutionary rate. Mol Biol Evol 28:2513–2520
49. Ha M, Lu J, Tian L, Ramachandran V, Kasschau KD, Chapman EJ, Carrington JC, Chen X, Wang XJ, Chen ZJ (2009) Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc Natl Acad Sci U S A 106: 17835–17840
50. Abrouk M, Zhang R, Murat F, Li A, Pont C, Mao L, Salse J (2012) Grass MiRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole-genome duplication. Plant Cell 24:1776–1792
51. Lehnert S, Van Loo P, Thilakarathne PJ, Marynen P, Verbeke G, Schuit FC (2009) Evidence for co-evolution between human miRNAs and Alu-repeats. PLoS One 4:e4456
52. Li J, Musso G, Zhang Z (2008) Preferential regulation of duplicated genes by miRNAs in mammals. Genome Biol 9:R132
53. D'Antonio M, Ciccarelli FD (2011) Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol 7:e1002029
54. Fernández A, Chen J (2009) Human capacitance to dosage imbalance: coping with inefficient selection. Genome Res 19:2185–2192
55. Felekkis K, Voskarides K, Dweep H, Sticht C, Gretz N, Deltas C (2011) Increased number of miRNA target sites in genes encoded in CNV regions. Evidence for an evolutionary genomic interaction. Mol Biol Evol 28:2421–2424
56. Dweep H, Gretz N, Felekkis K (2014) A schematic workflow of collecting information about the interaction of copy number variants and microRNAs with existing resources. Methods Mol Biol 1182:307–320
57. Henrichsen CN, Chaignat E, Reymond A (2009) Copy number variants, diseases and gene expression. Hum Mol Genet 18:R1–R8
58. Schuster-Bockler B, Conrad D, Bateman A (2010) Dosage sensitivity shapes the evolution of copy-number varied regions. PLoS One 5:e9474
59. Stranger BE, Forrest MS, Dunning M et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853
60. Wang RT, Ahn S, Park CC, Khan AH, Lange K, Smith DJ (2011) Effects of genome-wide copy number variation on expression in mammalian cells. BMC Genomics 16:562
61. Woodwark C, Bateman A (2011) The characterisation of three types of genes that overlie copy number variable regions. PLoS One 6:e14814
62. Karres JS, Hilgers V, Carrera I, Treisman J, Cohen SM (2007) The conserved miRNA miR-8 tunes atrophin levels to prevent neurodegeneration in Drosophila. Cell 131:136–145
63. Charroux B, Freeman M, Kerridge S, Baonza A (2006) Atrophin contributes to the negative regulation of epidermal growth factor receptor signaling in Drosophila. Dev Biol 29:278–290
64. Schnall-Levin M, Rissland OS, Johnston KW (2011) Unusually effective miRNA targeting within repeat-rich coding regions of mammalian mRNAs. Genome Res 21:1395–1403
65. Wu CI, Shen Y, Tang T (2009) Evolution under canalization and the dual roles of miRNAs: a hypothesis. Genome Res 19:734–743
66. Rutherford SL, Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. Nature 396:336–342
67. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP (2008) The impact of miRNAs on protein output. Nature 455:64–71

68. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N (2008) Widespread changes in protein synthesis induced by miRNAs. Nature 455:58–63
69. Gu X, Su Z, Huang Y (2009) Simultaneous expansions of miRNAs and protein-coding genes by gene/genome duplications in early vertebrates. J Exp Zool B Mol Dev Evol 312B: 164–170
70. Mukherji S, Ebert MS, Zheng GX, Tsang JS, Sharp PA, van Oudenaarden A (2011) MiRNAs can generate thresholds in target gene expression. Nat Genet 43:854–859

Chapter 2
MicroRNAs in Disease

Gregory Papagregoriou

The Study of MicroRNAs

Overview

MicroRNAs or miRNAs are a newly described class of short non-coding RNA molecules with the distinctive role of fine-tuning the expression of mRNAs in the living cells of all organisms [1]. Their targets are usually mRNA molecules that bear specific miRNA recognition sites on their 3′ UTRs, 5′ UTRs or coding regions; miRNAs bind onto their target sites in a Watson–Crick base pairing manner and eliminate mRNA translation (Fig. 2.1a) [2]. In humans, such post-transcriptional regulation of gene expression cannot be overlooked as more than half of all genes have evolutionary conserved miRNA target sites [3]. Consequently, miRNAs are prime regulators of all kinds of physiological cellular processes; hence faulty regulation of mRNA expression can lead to disease.

Target Prediction Algorithms

More than a handful of algorithms are available online and are elegantly designed to deliver predictions for miRNA binding sites on the 3′ UTR of protein-coding mRNAs, or to predict target sites for any miRNA sequence. Their numbers keep growing in order to serve the emerging demands of the scientific community for customization and credibility of results. Search variables usually include the

G. Papagregoriou, Ph.D. (✉)
Department of Biological Sciences, Molecular Medicine Research Center,
University of Cyprus, 75 Kallipoleos Str., 1678 Nicosia, Cyprus
e-mail: papagregoriou@ucy.ac.cy

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_2

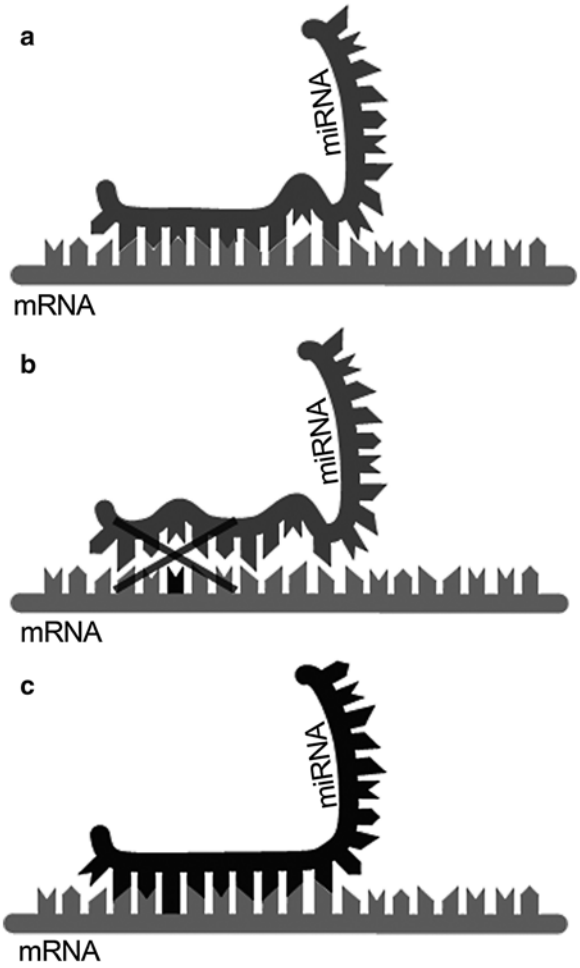


Fig. 2.1 Polymorphisms on miRNA target sequences (miRSNPs). Normal binding of miRNAs (**a**) can be diminished when target nucleotides corresponding to the miRNAs seed region are altered (**b**). In some cases, miRSNPs serve as gain-of-function mutations, when they create new target sites for different miRNAs (**c**)

seed-region length and miRNA/mRNA species, while each algorithm allows for further customization of search criteria depending on the prediction approach or even a cutoff p-value in correspondence to the statistical processing followed in each case. One must always have in mind the philosophy of predicting a miRNA target site on the 3′ UTR of an mRNA or elsewhere: a certain algorithm follows precise and predetermined criteria to predict such sites, which some are based solely on Watson–Crick complementarity between the miRNA and its target sites and other encompass more complicated approaches such as a learning algorithm or free energy values, therefore hits returned are destined to include false-positive results. Consequently, not all prediction results are valid and there is always a need for using a filtering approach in an effort to isolate the most useful and in a sense true miRNA–mRNA pairs. Filtering strategies can revolve around a wide or a narrow spectrum of criteria that are predominantly "*making sense*"; hence, both the miRNA and its potential mRNA targets should be expressed in the same tissue for example, or they are encountered in the same pathway or developmental stage. Moreover, all algorithms are enforced with a powerful statistical evaluation of prediction results, which in most cases can itself filter out correct pairs. In some studies, researchers repeat prediction analyses using a number of available algorithms and align prediction results to pick up only the pairs predicted by the most of them.

All algorithms predict target sites on 3′ UTR mRNA sequences, while some of them expand their predictions to include the gene's coding sequence, its 5′ UTR, sequences located upstream the transcription start points and even on the mitochondrial genome [4, 5]. The miRWalk algorithm works by "walking" on the gene sequence in a window of seven or more nucleotides and a miRNA is predicted to target a specific mRNA based only on seed region complementarity [4]. TargetScan on the other hand, is searching for the presence of conserved 8mer and 7mer sites that match the seed region of each miRNA, while it "ranks" a predicted miRNA target based on site length, type, context and accessibility [6]. The miRanda algorithm considers the miRNA sequence as input and searches a sequence dataset for potential target regions and successful predictions are made based on the alignment score and the minimum free energy of the miRNA bound to the potential target sequence [7]. The miRDB algorithm uses a completely different approach, as it is based on an SVM algorithm which is trained by a wiki database, in which users are able to input sequences of validated miRNA/mRNA couples [8], while the RNA22 algorithm uses a reverse approach as it first examines gene sequences for putative miRNA binding sites and then identifies a miRNA that could target an identified 3′ UTR site [9]. Predictions can be easily made by visiting the appropriate web location of each algorithm and placing a query about your miRNA or mRNA of choice. Results will be returned instantly and will depict the target site, its length, and a statistical score describing the likelihood this interaction is true based on the parameters on which each algorithm operates. The validity of prediction algorithms has been the number of many studies, all indicating an increased rate of false-positive results emerging from predictions, therefore a good bioinformatics analysis and filtering of predicted targets should be then supported by functional experiments [10].

Validation of miRNA Targets

As prediction algorithms can be a starting point in miRNA target discovery, direct interaction between a miRNA and an mRNA can only be valid when at least proven in vitro. Luciferase reporter constructs have been widely used as a straight-forward solution in studying direct binding of a miRNA on its target sequence. Albeit a useful and reliable tool, luciferase reporter constructs can only serve in examining one particular miRNA–mRNA target site and can be laborious at times. Target sites are introduced into the 3′-untranslated ending of the luciferase gene and plasmids are transfected into cell lines together with miRNA mimics or inhibitors (also called antagomirs). miRNA-analogous oligonucleotides are commercially available, relatively low in price and ready to use in cell cultures in various types; “mimics” that share the same sequence as a mature miRNA, “inhibitors” that are used to silence a specific miRNA, or “target site protectors” that prohibit miRNA binding onto its target sequence. Commercially available mimics and inhibitors are frequently chemically modified to LNAs (Locked Nucleic Acids) by the insertion of a 2′O-5′C Methyl-bridge, which helps the RNA oligonucleotide to keep an open conformation and to be thermally more stable than conventional oligos. (The reader is encouraged to see [11] and references within for a good presentation of currently used techniques in miRNA research and their limitations).

A good miRNA–mRNA couple can significantly reduce luciferase expression levels and in the same time a target site mutation, preferably at the nucleotides corresponding to the miRNA’s seed region, can abolish miRNA binding ability and therefore increase luciferase expression; this is a useful approach when investigating the effects of SNPs occurring at miRNA target sites.

As miRNAs are considered as post-transcriptional regulators, further investigation of their properties can include the determination of the protein levels of target mRNAs. This can be achieved by performing protein assays, i.e. western blots, amino acid stable isotope labelling or proteomics, after the overexpressing or knocking-down miRNAs in vitro. It has been found that specific miRNAs can regulate a restricted number of proteins, while changes in protein levels can sometimes be subtle [12]. Such effects are somehow expected, as protein levels are determined by a number of factors including mRNA transcription rate or protein degradation.

Identifying miRNAs in Tissues, Bodily Fluids and Exosomes

Isolation of miRNA species can be achieved by using readily available kits in the market developed by various companies. Specific applications require specific kits usually depending on the starting material, for example isolation of an enriched miRNA fraction from urine samples, serum or formalin-fixed paraffin-embedded (FFPE) tissues can be performed using appropriate commercial kit protocols. Alternatively, the TRIzol reagent can be used, although a biased loss of small RNAs with low GC content when processing a small number of cells with TRIzol has been reported [13]. Exosomes are small, 30–150 nm sized vesicles secreted by cells that contain miRNAs among other molecules; the isolation of miRNAs from exosomes can be achieved either by the use of Total exosome isolation reagents or by ultracentrifugation in sucrose gradients [14, 15]. Quantitation and quality assessment of enriched miRNA fractions can be performed using a microfluidics-based platform or an equivalent electrophoresis system, rather than a standard spectrophotometer due to their small size and reduced abundance compared to total RNA.

MicroRNAs can be detected using a number of methods, such as northern blots, real-time PCR or miRNA-specific probe hybridization. Although having limited sensitivity, northern blots are widely used for the identification of specific miRNAs and while their workflow is relatively simple, they require heavy optimization. A sample is let to run on an electrophoresis gel, which is consequently transferred to a porous membrane and miRNA-specific probes are let to hybridize onto targets while emitting fluorescence or radioactivity [16]. Relative quantification of cell-extracted miRNAs can be easily performed by real-time quantitative PCR (qRT-PCR). Following extraction, reverse transcription of miRNA species is performed in a two-step procedure: miRNA molecules are extended (3′ end) and single strand synthesis is completed with a universal primer [17, 18]. With the use of appropriate primers, miRNA sequences are detected in qRT-PCR and can be quantified by being compared to a number of small RNAs used as reference. Alternatively, reverse transcription can be performed using a stem-loop approach and miRNAs can be accurately detected via highly specific TaqMan probes [19]. In samples originating from exosomes or biofluids, quantitative analysis of miRNA expression can be difficult, as there is usually a lack of a stably-expressed small RNA to be used as a reference. For this purpose, a synthetic miRNA is usually spiked-in at a predetermined concentration to assist in downstream analyses [20–22].

In fresh or preserved tissue sections, miRNAs can be easily detected by in situ hybridization (ISH), where labelled probes with complementary sequences to target miRNAs are left to hybridize and reveal miRNA localization and differential expression [23]. Despite its wide use, ISH requires optimization and can sometimes be a laborious process. Nevertheless, when the probe affinity for its target is high, ISH can be used as a semiquantitative method of determining miRNA abundance in tissues and single cells. High resolution analysis of miRNA expression at single-cell level can be also performed by pairing fluorescent ISH with flow cytometry (flow-FISH), a technique able to give additional useful information on mRNAs or proteins of interest simultaneously [24, 25].

High-Throughput Methods in miRNA Research

The special nature of miRNA molecules makes their study a cumbersome matter; miRNAs are quite small in size and, unlike mRNAs, they do not share any common sequence features that ease their simultaneous isolation [26]. Tissue and cell miRNA profiling or disease biomarker discovery can be performed with the use of high-throughput methods such as commercially available microarray platforms or next-generation sequencing (NGS) of isolated miRNA fractions [27, 28]. Both techniques are considered acceptable, albeit approaching miRNA identification in a different manner. Microarrays identify fluorescently labelled miRNA cDNAs as they hybridize in complementary glass-immobilized probes, while NGS detects miRNAs by sequencing them; hence through NGS previously unidentified or novel miRNAs can be identified, while miRNA chip microarrays work with a predetermined range of miRNAs depending on the chip of choice. Differential expression of miRNAs on the other hand can be efficiently performed by both techniques, while small-scale study of specific miRNAs can be also performed using qRT-PCR [29]. Results are quite simple to read after they are further validated with qRT-PCR: certain miRNAs are expected to be either up- or down- regulated in a pathogenic tissue or cell type compared to controls, thus giving a sense of a pattern to characterize a disease.

Nevertheless, the elucidation of the exact biological meaning of any findings can at times be problematical and unfortunately in studies available, further functional investigation of findings is rarely performed.

Furthermore, microRNA targets can be also detected using high-throughput methods. Integration of high-throughput sequencing methods and protein immunoprecipitation, led to HITS-CLIP (High-throughput sequencing of RNA isolated with crosslinking immunoprecipitation) a robust method established by Chi et al. [10], which is exploited for the simultaneous isolation of miRNA–mRNA couples.

Antibodies raised against AGO are used for the immunoprecipitation of RNA-binding protein complexes from 254 nm UV-crosslinked samples, and at the same time the mRNAs bound on the miRNAs which these complexes accompany. An improved version of HITS-CLIP was developed by Hafner et al. [30] in an attempt to overcome technical limitations emerging from inefficient UV-crosslinking, named as Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP). In PAR-CLIP, cells are treated prior to crosslinking with 4-thiuridine, which is incorporated into targeted mRNAs to facilitate the precise binding position of the riboprotein complex by detecting thymidine to cytidine transitions. By PAR-CLIP, crosslinking efficiency was severely enhanced by irradiating cells with UV light and RNA recovery was dramatically improved. Moreover, Cross-linking ligation and sequencing of hybrids, or CLASH is a recently developed technique used to map miRNA–mRNA interactions by NGS [31]. In CLASH, cells that keep a stable expression a tagged AGO1 protein (PTG-AGO) are UV irradiated and lysed. PTH-AGO is then purified and samples are treated with RNAses that trim RNA–RNA duplexes, which are in turn ligated together and form chimeric miRNA–mRNA molecules that are eventually sequenced with NGS.

MicroRNAs Triggering Diseases

miRNA-Related Mutations as the Primary Cause of a Disease

A *miRSNP* (Fig. 2.1) can either be a single nucleotide change affecting the target region of a miRNA or its sequence at maturity. Such SNPs can effectively eliminate or weaken the binding of a miRNA to its target mRNA 3′ UTR (Fig. 2.1b) site and/or create a new binding site for a different miRNA (Fig. 2.1c); thus miRNAs can be considered as both primary or secondary players in disease development. In both cases, the protein levels of a targeted mRNA can be altered; at times, such changes can be phenotypically evident. Mutations in genes coding for miRNAs are considered as being quite rare; up to date only a small number of publications report such mutations that run in families in a Mendelian manner. Following the one-to-many mode of action, miRNAs with mutated seed regions lose the ability to target the range of mRNAs they usually aim at but inevitably gain novel targets. In some cases, mutations in the seed or other vital miRNA gene regions affect the abundance of the mature miRNA.

The frequency of SNPs in miRNA genes was thoroughly investigated by Gong et al. in 2011, who gathered all known SNPs from dbSNP v.132 that fell into such sequences. Notably, only 757 polymorphisms (SNPs and indels) out of 30 million in total were found to be located in 440 pre-miRNA regions, with 50 of them to be positioned in the seed regions of 41 miRNAs [32]. Such small numbers imply that there is a great possibility of a seed region sequence change to be a disease causing mutation rather than have no effect. Moreover, Gong et al. report that evolutionary conserved miRNAs and clustered miRNA genes tend to have less SNPs, a fact that can be attributed to the functional importance of certain miRNAs. The potential role for each SNP was calculated and recorded in the miRNASNP database (www.bioguo.org/miRNASNP).

The first work reporting seed region mutations was by Mencia et al., who identified two variants in miR-96, at positions 4 (+13G>A) and 5 (+14C>A) of its seed region, in a Spanish family with autosomal dominant non-syndromic hearing loss (ADNSHL) [33]. These mutations are responsible for the defected action of miR-96, as it fails to target and regulate five genes expressed in the inner ear; *AQP5*, *CELSR2*, *MYRIP*, *ODF2* and *RYK*. MiR-96 is expressed together with miR-182 and miR-183 as a multicistronic transcript in the mouse retina as well as in the inner ear. However, miR-96 seed region mutation carriers did not have an ocular phenotype, hence miR-96 is thought to target genes expressed in the ear rather than in the retina. In a different study, 882 ADNSHL patients from Italy were screened for miR-96 seed region mutations [34]. Interestingly, a novel mutation that segregated with the disease in one family was successfully identified, but was located outside the mature miRNA sequence. Being part of the pre-miRNA hairpin sequence, this mutation was found to effectively alter both mature miR-96 and miR-96* passenger miRNA biogenesis. Furthermore, Dorn et al. in 2012 described a similar mutation at the 3′ end (u17c) of miR-499 sequence that fell outside its seed region [35]. This mutation was identified while investigating a cohort of 2.606 individuals in search of genetic factors contributing in cardiomyopathy. By using luciferase reporter constructs and a mouse model, the authors identified a series of mRNAs that escaped miR-499 regulation possibly due to the c17 mutation.

Mutations in the miR-184 seed region have also been reported. MiR-184 was found to be abundantly expressed in the corneal and lens epithelia [36]. A single mutation at the fourth seed region nucleotide (c.57 C>U) of miR-184 was found to be associated with autosomal dominant familial keratoconus with early-onset anterior polar cataract in a Northern Irish family [37]. This mutation was identified after deep sequencing of a genomic locus indicated by linkage analysis in three generations of the family. The same mutation was also identified in a Spanish family with early onset cataract paired with various ocular abnormalities, while it has also been found in patients with EDICT syndrome (endothelial dystrophy, iris hypoplasia, congenital cataract and stromal thinning) [38, 39]. A different study identified two other mutations (+8C>A and +3A>G) in miR-184 in patients with isolated keratoconus that significantly repressed the expression of miR-184 [40]. Nevertheless, the exact mechanism explaining the pathogenesis of miR-184 seed region mutations remains elusive.

Mutations in miRNA Genes or Target Sites Contributing to Disease

Currently, there has been great interest in the discovery and functional characterization of miRSNPs located both on miRNA genes and miRNA target sites, as being contributors to a pathological phenotype. Such polymorphisms can act as phenotype modifiers by improving or exacerbating disease manifestation, or contributing to the risk of developing primary or secondary clinical states. Validated and predicted miRSNPs in human and mouse genes are recorded into four databases, Patrocles, dbSMR, PolymiRTS and MiRSNP [41–44]. Published data implicate miRSNPs in diseases in more ways than one; they can have a significant contribution to the pathogenesis of a disease, they can modify well characterized phenotypes of patients bearing a single mutation in a specific gene, they can regulate drug responses, or they can have no effect at all. Whatever the case may be, miRSNPs cannot be overlooked by modern geneticists. The role of miRSNPs will be analyzed below using some good examples from the current bibliography.

In diseases with monogenic inheritance, miRSNPs can have a significant effect. Evidence for such mechanism was shown for miR-24 when a point mutation that altered its binding to *SLITRK1* gene was identified in patients with Tourette syndrome [45]. Similarly, point mutations on *REEP1* which is a candidate gene for hereditary spastic paraplegia were found on the binding sites of two miRNAs (miR-140 and miR-691) [46, 47].

The role of miRSNPs as phenotype modifiers was demonstrated in CFHR5 nephropathy, where all patients found as far share an identical duplication of exons 2 and 3 in the *CFHR5* gene [48]. Although related, certain patients are clinically distinguishable with a portion of them rapidly progressing to mid-life end-stage renal failure requiring renal transplantation, and the rest having only some episodes of microscopic or macroscopic hematuria but with uncompromised renal function [49]. During the progression of the disease a genetic trigger channels the clinical fate of each patient towards a certain direction. The SNP rs13385 located on the 3′ UTR of the *HBEGF* gene and the target region of miR-1207-5p was reported to be associated with the severity of CFHR5 nephropathy in these patients, as the T-allele can eliminate the miRNA binding onto its target sequence [50].

Complex genetic traits are often the result of a joint action among a number of genetic and environmental factors that construct phenotypes, which are usually highly variable among patients. Impressively, a number of examples are available demonstrating the implication of the same common miRSNP in a number of different multifactorial phenotypes. For example, a miRSNP spotted on the miR-146a precursor (rs2910164 G>C) has been recently associated with susceptibility to leprosy [51]. When macrophage-like THP-1 cells were infected with live or irradiated strains of *Mycobacterium leprae*, live bacteria induced the expression of miR-146a, thus suggesting a pivotal role for this miRNA in disease progression. Consequently, C-allele carriers demonstrated a higher expression of miR-146a in nerves compared to patients with non-leprous neuropathies and this result was directly correlated with low levels of TNF recorded as a failure of the immune system to be effectively regulated. Impressively, the same SNP was also found to be associated with other multifactorial phenotypes, such as ischemic stroke, colorectal cancer survival [52], control of cell apoptosis, migration and growth in non-small cell lung cancer cells [53], or Hirschprung disease by eliminating *ROBO1* expression levels [54]. A different common SNP (rs11614913-C>T) on miR-162a2, was found to be related with increased susceptibility to esophageal squamous cell carcinomas [55], to the development of cardiovascular disease in patients with type 2 diabetes [56], but not Parkinson's disease [57]. Both SNPs on miR-146a and miR-162a2 are considered to be common polymorphisms, with Minor Allele Frequencies (MAF) 0.38 and 0.39 respectively, and together with rs71428439 (A/G, MAF 0.15) on miR-149, rs3746444 (A/G, MAF 0.18) on miR-499, rs895819 (T/C, MAF 0.36) on miR-27a stem loop, rs4938723 (T/C, MAF 0.31) on the Pri-miR-34b/c promoter form a team of miRSNPs that have been weakly or strongly associated with all kinds of diseases. However, only a small number of published works extend their findings in characterizing a functional relationship between the miRSNPs and relevant target genes to bridge gaps between genotypes and clinical phenotypes.

In a large number of publications, the role of miRSNPs in cancer susceptibility has been thoroughly evaluated [58]. Unfortunately, the functional characterization of such polymorphisms, as well as their comprehensive association to specific clinical features, is not the norm in most of them. Nevertheless, miRSNPs are proven to have an emerging role in cancer prognosis. A polymorphism on the 3′ UTR of *IGF-1R* gene limited the binding of miR-515-5p and increased the risk of developing breast cancer in subjects with *BRCA1* mutation [59]. In addition, increased susceptibility to breast cancer was attributed to two more SNPs identified on *TGFB1* and *XRCC1* which alter their expression levels as the target site of miR-187 and miR-183 is respectively interrupted [60]. In a Chinese lung cancer cohort, a SNP on the 3′ UTR of *CD133* was found to be significantly associated with a decreased risk in developing the disease, evidently by enhancing the binding of miR-135a/b to reduce CD133 levels [61].

Non-canonical miRNA Targeting Properties in Disease

As previously mentioned, miRNAs recognize and bind target sequences on the 3′ UTR of mRNAs waiting to be translated. In some rare cases, miRNAs were found to target mRNAs in other regions as well: the coding sequence and the 5′ UTR. Non-canonical miRNA targeting has been established by CLASH experiments, with 60 % of miRNAs to bind onto mRNAs with an irregular manner; mismatched nucleotides in the seed region, non-seed targeting and some of them binding to 5′ UTRs [31]. In addition, miRNAs were found to have similar efficiently in binding onto 3′ UTRs as they have when targeting 5′ UTRs in vitro to regulate the expression of mRNAs [62]. In some instances, miRNAs acting on the 5′ UTR of target mRNAs induce rather than repress their translation [63, 64]. Contrastingly, coding sequence targets are thought to be recognized by miRNAs less effectively compared to 3′ UTR target sites [65].

Akhtar et al. [66] presented evidence using luciferase reporter constructs that miR-602 and miR-608 target the sonic hedgehog (SHH) mRNA not on its 3′ UTR but on predicted target sites of the coding region. In osteoarthritis, *SHH* expression levels are elevated and eventually promote cartilage degradation. A potential mechanism explaining SHH upregulation involves the induction of SHH expression indirectly by IL-1β, through the direct suppression on the expression of miR-602 and miR-608 that repress SHH expression levels. In another study, the p53 inactivator MDM4 was found to have a miR-34a target site on its last 11th exon [67]. Moreover, a previously reported polymorphism rs79824231 located on this target site appeared to have the ability to disrupt miR-34a binding.

Functional miRNA target sites on the 5′ UTR are rarely found. Recently, miR-103a-3p was found to target and suppress the cancer related gene GPRC5A through two target sites on the gene's 5′ UTR in pancreatic cells [68]. Kim et al. [69] demonstrated a relationship between miR-34 family members, induced by p53, and the Axin2 mRNA in colorectal cancer. *Axin2* bears functional miR-34 target sites in both 5′ and 3′ UTRs that presumably act as "sponges" to negatively modulate miR-34 levels in cancer cells, which present elevated Axin2 and low miR-34 levels. Furthermore, a good example of non-canonical 5′ UTR miRNA binding is the ability of miR-122, a liver specific miRNA, to bind onto two 5′ UTR target sites of the Hepatitis C virus (HCV) genome and protects the UTR from host nucleolytic degradation to eventually promote its autonomous replication [70].

Emerging Pharmacogenomics Due to miRNA-Linked Genetic Variation

As expected, genomic variation related to miRNAs has also been implicated in patient response to administrated pharmaceutical therapy. In addition, certain drugs were found to formulate unique responses by interfering with miRNA expression levels. Such microRNAs have been described by a number of researchers and are frequently teamed under the term "Pharmaco-miRs". One can assume that mutated miRNA target sites on mRNAs engaged in pathways related to drug metabolism or absorption, can inevitably implicate miRNAs in drug responses as well [71]. In 2007 Mishra et al. presented evidence that miRSNPs can actually regulate drug responses. A polymorphism at the 3′ UTR of the dihydrofolate reductase (*DHFR*) gene, which was previously associated with upregulation of DFHR expression, was found to interrupt the conserved target site of miR-24 and caused the reported overexpression of DHFR [72]. As a result, elevated DHFR led to methotrexate resistance, a widely used chemotherapeutic agent. Polymorphisms in miRNA 3′ UTR target sites and/or miRNA genes have been associated with resistance to chemotherapy in patients suffering from various subtypes of cancer. The CREAM (Chemotherapy ResistancE-Associated MiRSNP) repository lists 150 such SNPs predicted to interfere with 1164 chemotherapy response compounds [73]. In breast cancer patients with estrogen receptor alpha expression, the acquired resistance to tamoxifen treatment has been associated with an elevated expression of miR-519a, which in turn targets a number of tumor-suppressor mRNAs to decrease the life expectancy of patients [74].

MicroRNAs in Developing Disease

Over the last few years many studies were designed to capture the effects of miRNA action in complex disease entities. Multifactorial diseases not only depend on the genetic load of an individual but are assisted towards overcoming a triggering threshold by lifestyle as well. MicroRNAs are considered as key regulators in disease development as they are implicated in all kinds of cellular processes and most importantly cell and tissue development and differentiation. Hence, the elucidation of miRNA signatures in such diseases, is postulated to unveil basic and advanced understanding of their pathology and progression and at the same time assist the development of tailored therapies per different patient. For this purpose, DICER knockout animal models, in vitro studies and expression assays have been recruited, as well as next generation high-throughput technologies.

MicroRNA signature of complex disease is a trending topic in scientific literature. Biomarkers are molecular signatures of a pathological state and can either be substances or molecules, which can be detected and measured in an objective way. MiRNAs have the specifications of being ideal and powerful biomarkers for non-invasive assays, as they can be measured in more than one ways, are abundantly expressed in all tissues and bodily fluids, are stable molecules and belong to a diverse and multitudinous family of non-coding RNAs.

miRNAs and Cancer

MicroRNA implication in cancer has been extensively studied in the past years and they have been found to be differentially expressed in a wide spectrum of malignant states. Being characterized as being both tumorigenic or tumor suppressive, miRNAs associated with cancer have been named as *Oncomirs* [75]; however, this characterization is only valid when a given miRNA is targeting an oncogene or a tumor suppressor gene. Deregulated miRNA expression under variable circumstances can alter their targeting potential against an mRNA, which in turn is associated with a specific type of cancer. In general, induction and progression of cancer is the orchestrated interaction and balance between tumor enhancers and suppressors and miRNAs seem to play a pivotal role in cancer development as more than half of miRNA encoding genes are found in genomic cancer hot-spots or at fragile chromosomal regions associated with translocations [76]. Tumor progression is marked by abnormal changes in cellular function and metabolism that lead to uncontrollable proliferation, resistance to apoptosis, escape from tumor suppressor action, induction of angiogenesis and eventually, invasion and metastasis; inevitably, miRNAs were found to be involved in all stages [77].

The first study implicating miRNAs in cancer by Calin et al. in 2002 demonstrated the involvement of miR-15a and miR-16-1 in B cell chronic lymphocytic leukemia. Their expression was found to be compromised by frequently observed deletions of the genes that encode for them in the 13q14 locus, which in turn is associated with the disease [78]. The miR-15a/miR-16 cluster of miRNA genes, encodes for miRNAs that are thought to be tumor suppressors under physiological states, by targeting BCL2 among other well described oncogenes [79]. The complexity of miRNA involvement in cancerous states has been widely observed. In some cases a specific miRNA has been given the properties of an oncomiR, while the same miRNA was found to act suppressively in other types of cancer. A good example is miR-125b; its levels drop in thyroid, ovarian and oral squamous cell carcinomas to halt cell proliferation and interfere with the progression of the cell-cycle, while in prostate cancer it was found to inhibit p53-dependent apoptosis of cancerous cells [80]. The p53 transcription factor is a direct regulator of miR-34a and miR-34b/c, while these miRNAs target and regulate p53 expression [81]. In chronic lymphocytic leukemia, miR-34a, miR-34b/c and *DAPK1* were found to be epigenetically inactivated by hypermethylation of their promoter to disrupt the tumor suppressive p53 pathway [82]. Hypermethylation of miR-34b/c is also considered as a potential diagnostic factor in Stage I non-small cell lung carcinoma [83]. In breast tumors, downregulation of miR-34a was significantly associated with metastasis [84].

In certain types of cancer miRNAs have a proven diagnostic and prognostic value, which has become of great significance in clinical practice. MicroRNA expression is evidently fluctuating in cancerous cells; affected tissues are distinguished by their expression potential of miRNAs. MiRNAs can be easily and efficiently isolated from formalin-fixed paraffin-embedded tissues, which is the starting material in most cases. Identification of miRNAs in biofluids, such as blood serum, saliva or urine, is also supporting the need of establishing non-invasive diagnostic and prognostic tests, as well as tumor classification tests.

The prognostic value of miRNAs has been investigated in many types of cancer, such as lung cancer, liver cancer, melanoma or prostate cancer. Inevitably, a number of specific miRNAs are recurrently found to be elevated or diminished in tissues or cell types studied. This fact can be explained by the potential role such miRNAs have in cancer development and their role as biomarkers cannot be overlooked. For example, in prostate cancer, miR-141 was found to be considerably elevated in the serum of patients with prostate cancer compared to controls and is considered as a prognostic marker with a 100 % specificity [85]. The same miRNA was also found to be increased in patients with ovarian cancer, while in metastatic colon cancer it was correlated with the levels of the carcinoembryogenic antigen (CEA) and poor prognosis [86, 87].

In a cohort of colon carcinoma patients, miR-21 was found to have a higher expression in adenomas as well as in patients with advance malignancy classification stage tumors. Survival of patients in the same study was also correlated with high miR-21 expression, as well as their response to therapy [88]. MiR-21 has also been found to be overly expressed in tumorous tissue from both breast and lungs [89, 90]. It is considered as a very important oncomiR and apoptosis suppressor and the therapeutic value of its inhibition was examined in breast cancer cells and mice with positive results [91]. In non-small cell lung carcinoma let-7 is not only considered as a primal tumor suppressor but also as a putative therapeutic agent. Poor prognosis of lung cancer in patients was directly correlated with compromised let-7 expression [92].

miRNAs and Diabetes

Diabetes mellitus (DM) is a complex disease affecting 347 million people of all ages worldwide. It is mainly characterized by elevated blood glucose levels and can be found in two forms depending on insulin availability or usage; pancreatic β-cells fail to produce insulin in Type 1 (T1D) patients, while Type 2 (T2D) patients are insulin resistant [93]. The role of miRNAs in diabetes starts with the control over pancreatic islet β-cell proliferation and function. In islet-specific Dicer1 mice knockouts, β-cells were completely absent and animals die soon after birth at P3 [94]. Beta-cell Dicer1 knockdown mice presented with a dramatic reduction in insulin production rates in isolated β-cells compared to wild-type animals [95]. Differentiation of insulin producing cells is a process mediated by a cross-talk between Neurogenin3 produced by endocrine progenitor cells and Hes1. Pancreatectomized mice presented a failure in β-cell regeneration from pancreatic pro-endocrine cells and defected post-transcriptional regulation of Neurogenin3 protein expression by miRNAs was found to be the cause; miR-15a, miR-15b, miR-16 and miR-195 were found to be highly expressed in treated mice and have predicted target sites on the Neurogenin3 transcript [96]. Beta-cell development and function is thought to be guarded specifically by miR-375. In vitro studies demonstrated an increase of insulin secretion after glucose stimulation in cells lacking miR-375, while miR-375 knockout mice presented with fasting hyperglycemia at the 12th week of life and β-cell mass was reduced responding to limited levels of proliferation [97, 98]. Using luciferase reporter constructs, miR-375 was found to have a conserved functional target site on the 3′ UTR of 3′-phosphoinositide-dependent protein kinase-1 (*PDK1*), an actively involved protein in insulin signaling and β-cell response in insulin demand through the phosphatidyloinositol 3 kinase (PI3-K) pathway [99]. Islet β-cell function is also regulated by miR-7a, which is regarded as a negative regulator of insulin granule exocytosis. In β-cell miR-7a2 knockout mice, insulin secretion was increased in response to high glucose levels, suggesting a higher tolerance to glucose [100].

Pancreatic β-cells sense glucose levels and respond by releasing insulin. Prolonged exposure of the pancreatic β-cell line MIN6 to high glucose levels initially induced miR-15a levels and eventually reduced them, in accordance with insulin production levels [101]. MiR-15a regulates insulin synthesis in an indirect manner, by targeting the mRNA of the uncoupling protein-2 gene (*UCP-2*), which codes for an important protein that monitors ATP generation triggered by glucose.

Insulin protein stability requires the polypyrimidine tract binding protein (PTB), which is in turn targeted by miR-133a. In glucose-treated human islets, miR-133 was found to be elevated and PTB biosynthesis was effectively reduced accompanied by a reduction in insulin synthesis as well [102]. In addition, insulin release is also regulated by miRNAs, with miR-124a targeting directly the exocytosis regulator Ras-related protein Rab27A, miR-96 and miR-9 increasing the expression levels of the Rab GTPase effector granuphilin, and miR-34a targeting the vesicle-associated membrane protein 2 (VAMP2) [103–105].

Tissue resistance to insulin leads to T2D and a number of studies demonstrate the involvement of miRNAs in this poorly understood process. A number of signaling pathways were found to be regulated by specific miRNAs in the liver, the adipose tissue and skeletal muscle, which uptake blood glucose in response to stimulation by endogenous or administrated insulin. Adipocyte development is facilitated by miR-143. Inhibition of miR-143 in pre-adipocytes reduced triglyceride accumulation by 75 %, while halted the expression of important genes such as *GLUT4*, *aP2*, *HSL* and *PPAR-γ2* [106]. The influence of miR-143 on adipogenesis was recently found to have a stage-specific role during their development, possibly by the direct regulation of MAP2K5 and consequently the MAPK signaling pathway [107]. In addition, insulin resistance in adipocytes is thought to be regulated by miR-320 via the direct regulation of the p85 PI3-K subunit, which in turn modulates the phosphorylation levels of Akt and Glut4 to assist downstream signaling pathways [108]. In obesity models, the pattern of miRNA expression in developing adipocytes appears to be paradoxically inversed [109].

The use of miRNAs as biomarkers in DM is also currently examined. In T1D non-obese mice, elevated miR-375 levels preceded the onset of diabetes by 2 weeks suggesting the use of this miRNA as a valid biomarker to indicate β-cell death and initiation of diabetes [110]. In T2D, a comprehensive evaluation of miRNAs in the plasma of patients, revealed the downregulation of miR-21, miR-24, miR-15a, miR-125, miR-191, miR-197, miR-223, miR-320 and miR-486 and the upregulation of miR-28-3p compared to healthy controls [111]. MiR-21 in particular, was found to be significantly upregulated in diabetic mice and correlated with the development of microalbuminuria and renal fibrosis and inflammation; soon after knocking down miR-21 in the same animals, renal symptoms ameliorated thus suggesting a potential role for miR-21 as a therapeutic agent for diabetic nephropathy [112]. Ethnic origin of T2D patients also appeared to play a role in circulating miRNA signatures. For instance, miR-144 was found to be significantly associated with T2D in Swedish patients, but not Iraqis, while miR-24 and miR-29b appeared to be consistently marking the disease in both populations [113].

miRNAs and Neurodegeneration

In neurodegenerative disorders, the physiological function or structure of neurons is gradually compromised leading to degeneration and inevitably cell death. Patients in most cases present with motor and/or cognitive decline depending on the impairment of specific brain regions, while clinicopathological symptoms occasionally overlap among different disease entities. Defected development of the central nervous system has been a common finding in Dicer knock-out animals, suggesting a pivotal role for miRNAs in neuronal differentiation and brain cortex size, while neurogenesis was found to be induced by a number of miRNAs, such as miR-9, miR-24, miR-125b and miR-128 [114, 115].

In sporadic Parkinson's disease (PD), mutations in *LRRK2* are considered as causative factors. LRRK2 protein levels are thought to be effectively regulated by miR-205 through a conserved binding site, while patients with sporadic PD demonstrated significantly reduced levels of miR-205, resulting in elevated LRRK2 levels and neurite outgrowth [116]. The role of LRRK2 in disease pathogenesis is still unclear, although its function as a kinase and GTPase implicates this protein in a number of molecular pathways possibly implicated with PD [117]. Nevertheless, in a drosophila model mutated *LRRK2* was found to interact with microRNA biogenesis through an RNA-independent association with the RISC complex [118]. Additionally, miR-133b was found to be reduced in the substantia nigra of PD patients and regulates the maturation and function of midbrain dopaminergic neurons possibly via Pitx3, which in turn modulates the expression level of miR-133b in a proposed negative feedback loop [119]. More genes involved in PD are also targeted by miRNAs, such as α-synuclein which was found to be regulated post-transcriptionally by miR-7 and miR-153 [120].

In developing Alzheimer's Disease (AD), specific miRNAs were found to regulate relevant genes. The β-amyloid precursor protein (APP) was found to be targeted by miR-106a and miR-502c [121], the tau protein is bound and regulated by miR-34a [122], and miR-98 targets IGF-1 [123]. Furthermore, exons 7 and 8 of APP are abnormally spliced in post-mitotic neurons of dicer conditionally knocked-out mice and miR-124 was found to be responsible for this effect assisted by its target gene *PTBP1* [124]. It appears that the abundance of miR-124 in neurons is concomitant with the occurrence of the neuronal APP isoform which lacks exons 7 and 8; hence, its absence promotes AD through the accumulation of non-neuronal APP. Furthermore, miR-9 was also found to be decreased in response to amyloid beta (Aβ) accumulation in primary neurons and is a direct regulator of *BACE1* expression, which in turn regulates APP cleavage [125, 126]. BACE1 mRNA is also targeted by miR-29 and miR-107, with the latter found to be downregulated in both AD and PD patients [119, 127, 128]. Moreover, in AD neuronal aging is promoted by an increase in miR-34 levels prior to the accumulation of Aβ in mice, possibly through its target gene *Bcl-2* [129].

The role of miR-146a in neurodegeneration has been explored by a number of studies. This miRNA is upregulated in brain regions affected in AD, while it is also induced by IL-1β, TNFα and Aβ42 peptides which in turn are pro-inflammatory cytokines triggering AD (reviewed in [130]). The same miRNA was also found to mark prion induced neurodegeneration [131]. The expression of miR-146a was found to be mainly induced by NF-κB in Toll/IL-1 receptors and represses the release of chemokines and IL-8 via a negative regulation of IL-1β as a part of a feedback loop to eventually regulate innate immune responses [132].

Identification of solid biomarkers in neurodegenerative disorders is still under great consideration by the scientific community. Examples of biomarkers in AD include miR-125b that has been proposed as a circulating biomarker with 68.3 % specificity and 80.8 % sensitivity when patients are compared to healthy controls [133]. Moreover, miR-384 which targets APP and BACE-1 was isolated from the cerebrospinal fluid of patients with mild cognitive impairment and Alzheimer's type dementia and was found to be significantly lower compared to controls [134]. In a recent meta-analysis of eight studies in search of the diagnostic validity of biomarkers in neurodegeneration, authors found contrasting results between different studies and suggest the usage of assays taking into account multiple miRNAs instead of a single species assay [135]. In other disorders such as multiple sclerosis, miR-21, miR-142-3p, miR-146a/b, miR-155 and miR-326 were found to be elevated in mononuclear blood cells and white matter brain regions in MS patients (reviewed in [136]). Also, in amyotrophic lateral sclerosis mice and patients the muscle-enriched miR-206 was found to be elevated in circulation and is proposed as a potential disease prognostic marker [137].

miRNAs and Cardiovascular Disease

The cardiovascular disease (CVD) group consists of a number of highly prevalent disorders that, together with cancer, are the leading causes of death in the western world [138]. Coronary artery disease presenting with acute myocardial infarction (MI), as well as essential hypertension, cardiac hypertrophy and atherosclerosis are the main categories of CVDs. Cardiac hypertrophy and failure are responses triggered by stress factors such as MI or hypertension, which in turn alter the hemodynamic environment and lead heart cells to undergo reprogramming in order to adapt. Consequently, heart cells regress to an expression potential that resembles a fetal type and cardiac cell specifically expressed miRNAs are thought to play a pivotal role in this procedure [139]. Dicer inactivation in cardiac cells of mice caused progressive dilated cardiomyopathy, accompanied by heart failure and eventually death after birth, while impaired expression of the dicer endonuclease was also identified in patients with dilated cardiomyopathy [140].

In the developing heart, miR-1, miR-133a/b and miR-208 are considered as basic players for regulating the expression of genes in cardiac muscles and therefore the heart's growth and function [141]. It has been demonstrated that at the onset of pressure-overload cardiac hypertrophy, miR-1 expression is repressed to cause downstream global changes in gene expression [142]. In Dahl hypertensive rats, silencing of miR-208a in the heart, prevented cardiac remodeling and myosin switching while increased rat survival [143].

MicroRNAs contribute in the development and stabilization of atherosclerotic plaques, with miR-92a to modulate plaque angiogenesis in mice [144], miR-21 to be stress-induced in the endothelium and modulates apoptosis [145], and miR-155 to regulate pro-inflammatory macrophages, act repressively on Bcl6 and enhance plaque formation [146], among others. In addition, miR-155, miR-145 and miR-126 are thought as potential candidates for atherosclerosis treatment [147]. Other factors conferring risk in atherosclerosis development, such as the high levels of low-density lipoprotein (LDL) and low levels of the high-density lipoprotein are regulated by miRNAs, with miR-122 to be responsible for increasing LDL levels and miR-33 upregulation to be associated with low HDL levels [148, 149]. Elevated LDL due to miR-122 upregulation is used as a control feature in miR-122 administrated antagomir therapy against Hepatitis C virus [148]. In an atherosclerotic mouse model, administration of miR-33 antagomirs increased HDL levels and plaque size appeared to be reduced and their stability increased [150].

In mice undergoing acute MI, members of the miR-29 family were found to be downregulated in the fibrotic heart region being adjacent to the infarct and to regulate the expression of fibrotic genes such as *COL1A1*, *COL1A2*, *COL3A1* and *FBN1* [151]. Cardiac fibrosis due to miR-29 might consequently lead to hypertension and it has been shown that adjustment of aerobic training habits in rats individuals helped in increasing miR-29c levels [152]. Moreover, miR-15 family members are found to be regulated in infarcted heart regions responding to ischemia-reperfusion injury in MI mice and pig models [153]. Therapeutic administration of miR-15 antagomirs, succeeded in sequestrating miR-15 tissue levels and reduced infarct size, while induced tissue remodeling.

In essential hypertension, human cytomegalovirus (HCMV)-encoded miRNA, hvmn-miR-UL112 was found to be differentially expressed between hypertensive patients and controls and it direct regulator of interferon regulatory factor 1 (IRF-1) mRNA, which in turn upregulates angiotensin II type3 receptor to deregulate blood pressure [154]. In pulmonary arterial hypertension (PAH), miR-204 was found to be significantly downregulated in vitro as a response to the aberrant expression of STAT3, which normally induces the transcription of miR-204 host gene [155]. Intratracheal delivery of miR-204 mimics in PAH affected rats presented with reduced arterial pulmonary pressure. Additionally, pulmonary vascular remodeling due to hypoxia in PAH has been associated with elevated miR-21 in mice [156].

The use of miRNAs as biomarkers in CVDs has also been extensively explored (reviewed in [157]). Myocardial damage is characterized by a release of miR-208b and miR-499 into circulation and the prognostic value of both miRNAs is under study [158]. In addition, the use of miR-133a as a biomarker for cardiomyocyte death in CVD patients has also been established [159]. Circulating levels of let-7b, miR-30a and miR-195 were identified in patients with developing MI and found to have up to 90 % sensitivity and 90 % specificity in discriminating patients from controls [160].

miRNAs and Renal Disease

Kidney development and growth is dependent on miRNAs. In mouse podocytes, the highly differentiated epithelial cells of the glomerulus, Dicer inactivation depleted foot processes and induced apoptosis, while animals developed albuminuria followed by glomerular sclerosis, tubulo-interstitial fibrosis, an abnormal appearance of the glomerular basement membrane, mesangial expansion, acute renal disease progression and eventually died after 6–8 weeks [161–163]. Impressively, proteins involved in podocyte function, such as nephrin and podocin, were found to be significantly decreased, while the major transcription factor that drives podocyte differentiation WT1 was found to be unaffected; hence miRNAs are believed to have limited influence in triggering podocyte differentiation but are essential in preserving podocyte function [163]. Like Dicer knockouts, clinical features of end-stage renal disease (ESRD) were recorded in Drosha-ablated mouse podocytes and animals presented with collapsing glomerulopathy, thus suggesting a pivotal role for miRNAs in podocyte function [164]. Microarray experiments revealed the specific enrichment of miR-192, miR-194, miR-204, miR-215 and miR-216 in the kidney, with miR-192 having a proven role in diabetic nephropathy (DN) [165, 166].

Diabetic patients developing nephropathy present proteinuria, thickening of the glomerular basement membrane (GBM), expansion of the mesangium and accumulation of the extracellular matrix, where laminin, fibronectin and collagens type I and IV fail to preserve GBM's normal structure [167]. In diabetic mice, miR-192 was found to be elevated in their glomeruli [168]. This miRNA can regulate E-box repressors expression, which in turn regulate *Col1a1* and *Col1a2* gene expression through TGF-β, leading to their accumulation. In addition, hyperglycemia seems to induce the abnormally high expression of miR-377 in mesangial cells, causing the targeted PAK1 and mnSOD mRNA downregulation and finally enhanced production of fibronectin [169]. Furthermore, in early diabetic nephropathy miRNA-21 was found to have reduced levels in db/db mice, while its over-expression paused mesangial cell proliferation and decreased the albumin excretion rate [170]. In podocytes cultured in high glucose levels, as well as in diabetic mice manifesting proteinuria, miR-195 expression was found to be elevated [171]. Furthermore, in Hypertensive Nephrosclerosis (HN), a disease where hypertension leads to arterial sclerosis which in turn causes glomerular sclerosis and hypertrophy, atrophy of the tubules and interstitial fibrosis, miR-200a and b, miR-141, miR-429, miR-205 and miR-192 were found in abundance and their levels correlated with the presence of proteinuria [172].

A frequent clinicopathological finding in glomerular disease is focal segmental glomerulosclerosis (FSGS), and urine miR-196a, miR-30a-5p and miR-490 were found to characterize patients with active FSGS compared to patients with FSGS in remission [173]. In mesangial glomerulonephritis, miR-21 and miR-124 were found to be upregulated in WKY rats [174]. Lupus nephritis (LN) manifests with mesangial glomerulonephritis, and in LN patients, miR-146a of glomerular origin was also found to be upregulated in both B6.MRLc1 mice and humans [175, 176]. In patients with IgA nephropathy, miR-200c was found to be downregulated with its levels of expression correlating with proteinuria, while miR-192, miR-141, miR-205 were upregulated, with miR-192 demonstrating an association with glomerulosclerosis and a decrease in glomerular filtration rate [177]. Furthermore, in a group of ESRD patients having received a kidney transplant, serum miR-181a, miR-483-5p and miR-557 were differentially expressed during the first 7 days following transplantation surgery, hence acting as factors predicting graft rejection [178].

In polycystic kidney disease (PKD), miRNAs were found to be important players in disease pathogenesis. Cyst development mechanisms in PKD are still poorly understood [179]. MiR-15a expression levels were found to be downregulated in both autosomal dominant and recessive PKD, which subsequently caused the upregulation of its target Cdc25A, a cell cycle regulator [180]. In a PKD rat model, differential expression of miRNAs demonstrated an increased expression of miR-21 and downregulation of miR-31, miR-164a and miR-125 [181]. Furthermore, the overexpression of the oncogenic miRNA cluster miR-17-92 in mice was found to be directly regulated with cyst growth, while its inactivation in a PKD mouse model delayed cyst development, as *Pkd1*, *Pkd2* and *Hnf-1b* gene transcripts have functional binding sites for both miR-17 and miR-92a [182].

MicroRNAs as Therapeutic Agents

The use of artificial miRNA molecules as therapeutic agents is currently under progressive development. Such molecules are modified oligonucleotides that can either sequestrate a specific miRNA to enhance the expression of its targets (antagomirs), or boost its levels to target disease promoting mRNAs (mimics). Based on their one-to-many way of function miRNA-like molecules can be used to treat diseases having more than one pathways affected, but on the other hand limit the specificity of a putative therapeutic approach [183]. Hence, good therapeutic candidates can be miRNAs that have a well-documented way of action and are ideally tissue specific. Moreover, drug delivery can be problematic and challenging in some cases (reviewed in [184]).

As mentioned earlier, in a work by Joplin et al. [70] miR-122 was found to target two specific sites at the 5′ UTR of the HCV genome to stabilize and protect it from nucleolytic degradation, thus assisting in hepatitis C propagation [185]. HCV infection is the primary cause of liver disease and affects more than 170 million people worldwide; hence the development of effectual therapeutic approaches is more than necessary and the exploitation of miR-122 properties over HCV will hopefully give rise to the first approved miRNA-related drug [186]. To develop this *miravirsen* drug, a series of anti-miR-122 oligonucleotides targeting the miRNA's 5′ end sequence were tested in order to block its seed region, and finally a highly stable sequence specific 15-mer LNA-modified oligo (SPC3649) demonstrated strong inhibitory effects on miR-122 function in hepatocytes of both mice and African green monkeys, even at low concentrations [148, 187]. When this miravirsen was administrated in HCV infected chimpanzees, they presented a significant reduction in traceable HCV, while HCV levels continued to drop even 2 weeks after the drug administration was concluded, thus rendering the miravirsen a more comprehensive therapeutic agent compared to other approaches [188]. Currently, this miravirsen is in Phase 2a clinical trials with very encouraging results. Clinical trials exhibit a dose-dependent and persistent reduction of HCV levels, while no side-effects were recorded [189]. Impressively, 14 weeks after concluding the drug administration four out of nine patients that received a high dose (7 mg/kg) of miravirsen, the viral RNA was undetectable, depicting a lack of resistance of the virus against the drug.

Besides miR-122, a small number of other miRNAs are also thought as potential therapeutic candidates. For example, miR-92a was found to control angiogenesis and its inhibition through an antagomir in mouse models presenting with limb ischemia and myocardial infarction, accelerated the recovery of affected tissues and promoted angiogenesis as it naturally targets relevant genes [144]. Additionally, the members of miR-15 family are found to be implicated in cell survival and regulate cell cycle progression and are upregulated in infarcted regions of mice and pigs in response to injury caused by ischemia [153]. Administration of modified LNAs in animals sequestrated the expression levels of miR-15 family members and increased the viability of cardiomyocytes after hypoxia, while cardiac function was enhanced.

Another important miRNA drug under development is MRX34, a double stranded RNA molecule that mimics miR-34 and is currently in Phase 1 clinical trials. As mentioned above, miR-34 family expression is promoted by p53 and is found to be downregulated in various types of cancer, as it is considered an important tumor suppressor (reviewed in [183]). The miR-34 family consists of three miRNAs that share the same seed-region, namely miR-34a, b and c, with miR-34a to be the most abundant. Tumor suppression in vivo was succeeded in animals with xenografts that promoted prostate cancer [190], lymphoma [191], non-small cell lung cancer [192] and others, which had miR-34 delivered intravenously or directly with intratumoral injections. Phase 1 trials of miR-34 replacement are ongoing and started with a debate on the drug delivery approach to be used. After exploring a number of available solutions, an ionizable liposome (NOV340—SMARTICLES) was selected to carry the miR-34 oligonucleotide based on its use in mouse models, the miRNA bio-distribution and the vehicle's safety [193]. When this liposome is released in biofluids with neutral pH it gains a slightly anionic character that prevents it from having non-specific interactions with negatively charged cellular membranes, while it becomes cationic in tumor regions where the environment has lower pH and becomes active. This specific vehicle is preferably delivered to the liver, thus liver cancer was proposed as the disease model in this case.

Conclusions

Without doubt, miRNA involvement in diseases is a trending matter in the literature. From causing a disease to modifying complex genotypes, miRNAs seem to be implicated in every aspect of a cell's effort to develop, differentiate, and lead a healthy living or program its death. Although a number of different approaches both in vitro and in vivo have been developed as a response to the growing needs of the scientific community to study miRNAs, the need to discover robust prognostic and diagnostic miRNA markers is essential. In addition, miRNA bioinformatics have been greatly used as the starting point in deciphering miRNA–mRNA target pairings and have become more and more efficient through time. Hopefully the development of novel high-throughput technique in the years to come will facilitate the comprehensive functional characterization of miRNAs in disease; the knowledge emerging from endless lists of differentially expressed miRNAs in cells or tissues is of little help in understanding their true biological meaning.

Conclusively, miRNAs have what it takes to become the next generation of therapeutic agents. Taking miRNA research from the bench to the bedside, it is indeed exciting that the first two miRNA drugs, miravirsen and miR-34 replacement, are already in clinical trials. What remains is the development of more drugs and why not, tailored therapies for patients.

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116(2): 281–297
2. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136(2): 215–233
3. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19(1):92–105
4. Dweep H, Gretz N, Sticht C (2014) miRWalk database for miRNA-target interactions. Methods Mol Biol 1182:289–305
5. Kumar A, Wong AK, Tizard ML, Moore RJ, Lefevre C (2012) miRNA_Targets: a database for miRNA target predictions in coding and non-coding regions of mRNAs. Genomics 100(6):352–356
6. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120(1):15–20
7. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS (2003) MicroRNA targets in Drosophila. Genome Biol 5(1):R1
8. Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 14(6):1012–1017
9. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM et al (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 126(6):1203–1217
10. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 460(7254):479–486
11. Zhang B, Pan X, Cobb GP, Anderson TA (2007) microRNAs as oncogenes and tumor suppressors. Dev Biol 302(1):1–12
12. Vinther J, Hedegaard MM, Gardner PP, Andersen JS, Arctander P (2006) Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. Nucleic Acids Res 34(16):e107
13. Kim YK, Yeo J, Kim B, Ha M, Kim VN (2012) Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. Mol Cell 46(6):893–895
14. Zeringer E, Li M, Barta T, Schageman J, Pedersen KW, Neurauter A et al (2013) Methods for the extraction and RNA profiling of exosomes. World J Methodol 3(1):11–18
15. Schageman J, Zeringer E, Li M, Barta T, Lea K, Gu J et al (2013) The complete exosome workflow solution: from isolation to characterization of RNA cargo. Biomed Res Int 2013:253957
16. Rio DC (2014) Northern blots for small RNAs and microRNAs. Cold Spring Harb Protoc 2014(7):793–797

17. Raymond CK, Roberts BS, Garrett-Engele P, Lim LP, Johnson JM (2005) Simple, quantitative primer-extension PCR assay for direct monitoring of microRNAs and short-interfering RNAs. RNA 11(11):1737–1744
18. Ro S, Park C, Jin J, Sanders KM, Yan W (2006) A PCR-based method for detection and quantification of small RNAs. Biochem Biophys Res Commun 351(3):756–763
19. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT et al (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33(20):e179
20. Li Y, Kowdley KV (2012) Method for microRNA isolation from clinical serum samples. Anal Biochem 431(1):69–75
21. McAlexander MA, Phillips MJ, Witwer KW (2013) Comparison of methods for miRNA extraction from plasma and quantitative recovery of RNA from cerebrospinal fluid. Front Genet 4:83
22. Turchinovich A, Weiz L, Langheinz A, Burwinkel B (2011) Characterization of extracellular circulating microRNA. Nucleic Acids Res 39(16):7223–7233
23. Nuovo GJ (2008) In situ detection of precursor and mature microRNAs in paraffin embedded, formalin fixed tissues and cell preparations. Methods 44(1):39–46
24. Wu M, Piccini M, Koh CY, Lam KS, Singh AK (2013) Single cell microRNA analysis using microfluidic flow cytometry. PLoS One 8(1):e55044
25. Porichis F, Hart MG, Griesbeck M, Everett HL, Hassan M, Baxter AE et al (2014) High-throughput detection of miRNAs and gene-specific mRNA at the single-cell level by flow cytometry. Nat Commun 5:5641
26. de Planell-Saguer M, Rodicio MC (2011) Analytical aspects of microRNA in diagnostics: a review. Anal Chim Acta 699(2):134–152
27. Yin JQ, Zhao RC, Morris KV (2008) Profiling microRNA expression with microarrays. Trends Biotechnol 26(2):70–76
28. Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C et al (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods 44(1): 3–12
29. Zollner H, Hahn SA, Maghnouj A (2014) Quantitative RT-PCR specific for precursor and mature miRNAs. Methods Mol Biol 1095:121–134
30. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P et al (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141(1):129–141
31. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 153(3):654–665
32. Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G et al (2012) Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum Mutat 33(1):254–263
33. Mencia A, Modamio-Hoybjor S, Redshaw N, Morin M, Mayo-Merino F, Olavarrieta L et al (2009) Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat Genet 41(5):609–613
34. Solda G, Robusto M, Primignani P, Castorina P, Benzoni E, Cesarani A et al (2012) A novel mutation within the MIR96 gene causes non-syndromic inherited hearing loss in an Italian family by altering pre-miRNA processing. Hum Mol Genet 21(3):577–585
35. Dorn GW 2nd, Matkovich SJ, Eschenbacher WH, Zhang Y (2012) A human 3′ miR-499 mutation alters cardiac mRNA targeting and function. Circ Res 110(7):958–967
36. Ryan DG, Oliveira-Fernandes M, Lavker RM (2006) MicroRNAs of the mammalian eye display distinct and overlapping tissue specificity. Mol Vis 12:1175–1184
37. Hughes AE, Bradley DT, Campbell M, Lechner J, Dash DP, Simpson DA et al (2011) Mutation altering the miR-184 seed region causes familial keratoconus with cataract. Am J Hum Genet 89(5):628–633
38. Iliff BW, Riazuddin SA, Gottsch JD (2012) A single-base substitution in the seed region of miR-184 causes EDICT syndrome. Invest Ophthalmol Vis Sci 53(1):348–353

39. Bykhovskaya Y, Caiado Canedo AL, Wright KW, Rabinowitz YS (2013) C.57 C > T Mutation in MIR 184 is Responsible for Congenital Cataracts and Corneal Abnormalities in a Five-generation Family from Galicia, Spain. Ophthalmic Genet DOI:10.3109/13816810.2013.848908
40. Lechner J, Bae HA, Guduric-Fuchs J, Rice A, Govindarajan G, Siddiqui S et al (2013) Mutational analysis of MIR184 in sporadic keratoconus and myopia. Invest Ophthalmol Vis Sci 54(8):5266–5272
41. Georges M, Clop A, Marcq F, Takeda H, Pirottin D, Hiard S et al (2006) Polymorphic microRNA-target interactions: a novel source of phenotypic variation. Cold Spring Harb Symp Quant Biol 71:343–350
42. Hariharan M, Scaria V, Brahmachari SK (2009) dbSMR: a novel resource of genome-wide SNPs affecting microRNA mediated regulation. BMC Bioinformatics 10:108
43. Bao L, Zhou M, Wu L, Lu L, Goldowitz D, Williams RW et al (2007) PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. Nucleic Acids Res 35(Database issue):D51–D54
44. Liu C, Zhang F, Li T, Lu M, Wang L, Yue W et al (2012) MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics 13:661
45. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM et al (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 310(5746):317–320
46. Beetz C, Schule R, Deconinck T, Tran-Viet KN, Zhu H, Kremer BP et al (2008) REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. Brain 131(Pt 4):1078–1086
47. Zuchner S, Wang G, Tran-Viet KN, Nance MA, Gaskell PC, Vance JM et al (2006) Mutations in the novel mitochondrial protein REEP1 cause hereditary spastic paraplegia type 31. Am J Hum Genet 79(2):365–369
48. Gale DP, de Jorge EG, Cook HT, Martinez-Barricarte R, Hadjisavvas A, McLean AG et al (2010) Identification of a mutation in complement factor H-related protein 5 in patients of Cypriot origin with glomerulonephritis. Lancet 376(9743):794–801
49. Athanasiou Y, Voskarides K, Gale DP, Damianou L, Patsias C, Zavros M et al (2011) Familial C3 glomerulopathy associated with CFHR5 mutations: clinical characteristics of 91 patients in 16 pedigrees. Clin J Am Soc Nephrol 6(6):1436–1446
50. Papagregoriou G, Erguler K, Dweep H, Voskarides K, Koupepidou P, Athanasiou Y et al (2012) A miR-1207-5p binding site polymorphism abolishes regulation of HBEGF and is associated with disease severity in CFHR5 nephropathy. PLoS One 7(2):e31021
51. Cezar-de-Mello PF, Toledo-Pinto TG, Marques CS, Arnez LE, Cardoso CC, Guerreiro LT et al (2014) Pre-miR-146a (rs2910164 G>C) single nucleotide polymorphism is genetically and functionally associated with leprosy. PLoS Negl Trop Dis 8(9):e3099
52. Chae YS, Kim JG, Lee SJ, Kang BW, Lee YJ, Park JY et al (2013) A miR-146a polymorphism (rs2910164) predicts risk of and survival from colorectal cancer. Anticancer Res 33(8):3233–3239
53. Chen G, Umelo IA, Lv S, Teugels E, Fostier K, Kronenberger P et al (2013) miR-146a inhibits cell growth, cell migration and induces apoptosis in non-small cell lung cancer cells. PLoS One 8(3):e60317
54. Zhu H, Cai P, Zhu D, Xu C, Li H, Tang J et al (2014) A common polymorphism in pre-miR-146a underlies Hirschsprung disease risk in Han Chinese. Exp Mol Pathol 97(3):511–514
55. Wang N, Li Y, Zhou RM, Wang GY, Wang CM, Chen ZF et al (2014) Hsa-miR-196a2 functional SNP is associated with the risk of ESCC in individuals under 60 years old. Biomarkers 19(1):43–48
56. Buraczynska M, Zukowski P, Wacinski P, Ksiazek K, Zaluska W (2014) Polymorphism in microRNA-196a2 contributes to the risk of cardiovascular disease in type 2 diabetes patients. J Diabetes Complications 28(5):617–620
57. Haixia D, Hairong D, Weixian C, Min Y, Qiang W, Hang X (2012) Lack of association of polymorphism in miRNA-196a2 with Parkinson's disease risk in a Chinese population. Neurosci Lett 514(2):194–197

58. Ryan BM, Robles AI, Harris CC (2010) Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer 10(6):389–402
59. Gilam A, Edry L, Mamluk-Morag E, Bar-Ilan D, Avivi C, Golan D et al (2013) Involvement of IGF-1R regulation by miR-515-5p modifies breast cancer risk among BRCA1 carriers. Breast Cancer Res Treat 138(3):753–760
60. Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M et al (2010) Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. Cancer Res 70(7):2789–2798
61. Cheng M, Yang L, Yang R, Yang X, Deng J, Yu B et al (2013) A microRNA-135a/b binding polymorphism in CD133 confers decreased risk and favorable prognosis of lung cancer in Chinese by reducing CD133 expression. Carcinogenesis 34(10):2292–2299
62. Lytle JR, Yario TA, Steitz JA (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. Proc Natl Acad Sci U S A 104(23):9667–9672
63. Tsai NP, Lin YL, Wei LN (2009) MicroRNA mir-346 targets the 5′-untranslated region of receptor-interacting protein 140 (RIP140) mRNA and up-regulates its protein expression. Biochem J 424(3):411–418
64. Orom UA, Nielsen FC, Lund AH (2008) MicroRNA-10a binds the 5′ UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell 30(4):460–471
65. Forman JJ, Coller HA (2010) The code within the code: microRNAs target coding regions. Cell Cycle 9(8):1533–1541
66. Akhtar N, Makki MS, Haqqi TM (2015) MicroRNA-602 and microRNA-608 regulate sonic hedgehog expression via target sites in the coding region in human chondrocytes. Arthritis Rheumatol 67(2):423–434
67. Mandke P, Wyatt N, Fraser J, Bates B, Berberich SJ, Markey MP (2012) MicroRNA-34a modulates MDM4 expression via a target site in the open reading frame. PLoS One 7(8):e42034
68. Zhou H, Rigoutsos I (2014) MiR-103a-3p targets the 5′ UTR of GPRC5A in pancreatic cells. RNA 20(9):1431–1439
69. Kim NH, Cha YH, Kang SE, Lee Y, Lee I, Cha SY et al (2013) p53 regulates nuclear GSK-3 levels through miR-34-mediated Axin2 suppression in colorectal cancer cells. Cell Cycle 12(10):1578–1587
70. Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P (2005) Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science 309(5740):1577–1581
71. Mishra PJ, Banerjee D, Bertino JR (2008) MiRSNPs or MiR-polymorphisms, new players in microRNA mediated regulation of the cell: Introducing microRNA pharmacogenomics. Cell Cycle 7(7):853–858
72. Mishra PJ, Humeniuk R, Longo-Sorbello GS, Banerjee D, Bertino JR (2007) A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. Proc Natl Acad Sci U S A 104(33):13513–13518
73. Dai E, Lv Y, Meng F, Yu X, Zhang Y, Wang S et al (2014) CREAM: a database for chemotherapy resistance-associated miRSNP. Cell Death Dis 5:e1272
74. Ward A, Shukla K, Balwierz A, Soons Z, Konig R, Sahin O et al (2014) MicroRNA-519a is a novel oncomir conferring tamoxifen resistance by targeting a network of tumour-suppressor genes in ER+ breast cancer. J Pathol 233(4):368–379
75. Esquela-Kerscher A, Slack FJ (2006) Oncomirs – microRNAs with a role in cancer. Nat Rev Cancer 6(4):259–269
76. Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S et al (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A 101(9):2999–3004
77. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674
78. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E et al (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 99(24):15524–15529

79. Aqeilan RI, Calin GA, Croce CM (2010) miR-15a and miR-16-1 in cancer: discovery, function and future perspectives. Cell Death Differ 17(2):215–220
80. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA (2011) MicroRNAs in body fluids--the mix of hormones and biomarkers. Nat Rev Clin Oncol 8(8):467–477
81. Hermeking H (2010) The miR-34 family in cancer and apoptosis. Cell Death Differ 17(2):193–199
82. Wang LQ, Kwong YL, Wong KF, Kho CS, Jin DY, Tse E et al (2014) Epigenetic inactivation of mir-34b/c in addition to mir-34a and DAPK1 in chronic lymphocytic leukemia. J Transl Med 12:52
83. Wang Z, Chen Z, Gao Y, Li N, Li B, Tan F et al (2011) DNA hypermethylation of microRNA-34b/c has prognostic value for stage non-small cell lung cancer. Cancer Biol Ther 11(5):490–496
84. Javeri A, Ghaffarpour M, Taha MF, Houshmand M (2013) Downregulation of miR-34a in breast tumors is not associated with either p53 mutations or promoter hypermethylation while it correlates with metastasis. Med Oncol 30(1):413
85. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL et al (2008) Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A 105(30):10513–10518
86. Taylor DD, Gercel-Taylor C (2008) MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. Gynecol Oncol 110(1):13–21
87. Cheng H, Zhang L, Cogdell DE, Zheng H, Schetter AJ, Nykter M et al (2011) Circulating plasma MiR-141 is a novel biomarker for metastatic colon cancer and predicts poor prognosis. PLoS One 6(3):e17745
88. Schetter AJ, Leung SY, Sohn JJ, Zanetti KA, Bowman ED, Yanaihara N et al (2008) MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. JAMA 299(4):425–436
89. Iorio MV, Casalini P, Tagliabue E, Menard S, Croce CM (2008) MicroRNA profiling as a tool to understand prognosis, therapy response and resistance in breast cancer. Eur J Cancer 44(18):2753–2759
90. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M et al (2006) Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 9(3):189–198
91. Si ML, Zhu S, Wu H, Lu Z, Wu F, Mo YY (2007) miR-21-mediated tumor growth. Oncogene 26(19):2799–2803
92. Markou A, Liang Y, Lianidou E (2011) Prognostic, therapeutic and diagnostic potential of microRNAs in non-small cell lung cancer. Clin Chem Lab Med 49(10):1591–1603
93. Chen H, Lan HY, Roukos DH, Cho WC (2014) Application of microRNAs in diabetes mellitus. J Endocrinol 222(1):R1–R10
94. Lynn FC, Skewes-Cox P, Kosaka Y, McManus MT, Harfe BD, German MS (2007) MicroRNA expression is required for pancreatic islet cell genesis in the mouse. Diabetes 56(12):2938–2945
95. Kalis M, Bolmeson C, Esguerra JL, Gupta S, Edlund A, Tormo-Badia N et al (2011) Beta-cell specific deletion of Dicer1 leads to defective insulin secretion and diabetes mellitus. PLoS One 6(12):e29166
96. Joglekar MV, Parekh VS, Mehta S, Bhonde RR, Hardikar AA (2007) MicroRNA profiling of developing and regenerating pancreas reveal post-transcriptional regulation of neurogenin3. Dev Biol 311(2):603–612
97. Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE et al (2004) A pancreatic islet-specific microRNA regulates insulin secretion. Nature 432(7014):226–230
98. Poy MN, Hausser J, Trajkovski M, Braun M, Collins S, Rorsman P et al (2009) miR-375 maintains normal pancreatic alpha- and beta-cell mass. Proc Natl Acad Sci U S A 106(14):5813–5818

99. El Ouaamari A, Baroukh N, Martens GA, Lebrun P, Pipeleers D, van Obberghen E (2008) miR-375 targets 3′-phosphoinositide-dependent protein kinase-1 and regulates glucose-induced biological responses in pancreatic beta-cells. Diabetes 57(10):2708–2717
100. Latreille M, Hausser J, Stutzer I, Zhang Q, Hastoy B, Gargani S et al (2014) MicroRNA-7a regulates pancreatic beta cell function. J Clin Invest 124(6):2722–2735
101. Sun LL, Jiang BG, Li WT, Zou JJ, Shi YQ, Liu ZM (2011) MicroRNA-15a positively regulates insulin synthesis by inhibiting uncoupling protein-2 expression. Diabetes Res Clin Pract 91(1):94–100
102. Fred RG, Bang-Berthelsen CH, Mandrup-Poulsen T, Grunnet LG, Welsh N (2010) High glucose suppresses human islet insulin biosynthesis by inducing miR-133a leading to decreased polypyrimidine tract binding protein-expression. PLoS One 5(5):e10843
103. Lovis P, Gattesco S, Regazzi R (2008) Regulation of the expression of components of the exocytotic machinery of insulin-secreting cells by microRNAs. Biol Chem 389(3):305–312
104. Plaisance V, Abderrahmani A, Perret-Menoud V, Jacquemin P, Lemaigre F, Regazzi R (2006) MicroRNA-9 controls the expression of Granuphilin/Slp4 and the secretory response of insulin-producing cells. J Biol Chem 281(37):26932–26942
105. Lovis P, Roggli E, Laybutt DR, Gattesco S, Yang JY, Widmann C et al (2008) Alterations in microRNA expression contribute to fatty acid-induced pancreatic beta-cell dysfunction. Diabetes 57(10):2728–2736
106. Esau C, Kang X, Peralta E, Hanson E, Marcusson EG, Ravichandran LV et al (2004) MicroRNA-143 regulates adipocyte differentiation. J Biol Chem 279(50):52361–52365
107. Chen L, Hou J, Ye L, Chen Y, Cui J, Tian W et al (2014) MicroRNA-143 regulates adipogenesis by modulating the MAP2K5-ERK5 signaling. Sci Rep 4:3819
108. Ling HY, Ou HS, Feng SD, Zhang XY, Tuo QH, Chen LX et al (2009) CHANGES IN microRNA (miR) profile and effects of miR-320 in insulin-resistant 3T3-L1 adipocytes. Clin Exp Pharmacol Physiol 36(9):e32–e39
109. Xie H, Lim B, Lodish HF (2009) MicroRNAs induced during adipogenesis that accelerate fat cell development are downregulated in obesity. Diabetes 58(5):1050–1057
110. Erener S, Mojibian M, Fox JK, Denroche HC, Kieffer TJ (2013) Circulating miR-375 as a biomarker of beta-cell death and diabetes in mice. Endocrinology 154(2):603–608
111. Zampetaki A, Kiechl S, Drozdov I, Willeit P, Mayr U, Prokopi M et al (2010) Plasma microRNA profiling reveals loss of endothelial miR-126 and other microRNAs in type 2 diabetes. Circ Res 107(6):810–817
112. Zhong X, Chung AC, Chen HY, Dong Y, Meng XM, Li R et al (2013) miR-21 is a key therapeutic target for renal injury in a mouse model of type 2 diabetes. Diabetologia 56(3):663–674
113. Wang X, Sundquist J, Zoller B, Memon AA, Palmer K, Sundquist K et al (2014) Determination of 14 circulating microRNAs in Swedes and Iraqis with and without diabetes mellitus type 2. PLoS One 9(1):e86792
114. Krichevsky AM, Sonntag KC, Isacson O, Kosik KS (2006) Specific microRNAs modulate embryonic stem cell-derived neurogenesis. Stem Cells 24(4):857–864
115. Kawase-Koga Y, Low R, Otaegi G, Pollock A, Deng H, Eisenhaber F et al (2010) RNAase-III enzyme Dicer maintains signaling pathways for differentiation and survival in mouse cortical neural stem cells. J Cell Sci 123(Pt 4):586–594
116. Cho HJ, Liu G, Jin SM, Parisiadou L, Xie C, Yu J et al (2013) MicroRNA-205 regulates the expression of Parkinson’s disease-related leucine-rich repeat kinase 2 protein. Hum Mol Genet 22(3):608–620
117. Esteves AR, Swerdlow RH, Cardoso SM (2014) LRRK2, a puzzling protein: insights into Parkinson’s disease pathogenesis. Exp Neurol 261:206–216
118. Gehrke S, Imai Y, Sokol N, Lu B (2010) Pathogenic LRRK2 negatively regulates microRNA-mediated translational repression. Nature 466(7306):637–641
119. Kim J, Inoue K, Ishii J, Vanti WB, Voronov SV, Murchison E et al (2007) A microRNA feedback circuit in midbrain dopamine neurons. Science 317(5842):1220–1224

120. Doxakis E (2010) Post-transcriptional regulation of alpha-synuclein expression by mir-7 and mir-153. J Biol Chem 285(17):12726–12734
121. Patel N, Hoang D, Miller N, Ansaloni S, Huang Q, Rogers JT et al (2008) MicroRNAs can regulate human APP levels. Mol Neurodegener 3:10
122. Dickson JR, Kruse C, Montagna DR, Finsen B, Wolfe MS (2013) Alternative polyadenylation and miR-34 family members regulate tau expression. J Neurochem 127(6):739–749
123. Hu YK, Wang X, Li L, Du YH, Ye HT, Li CY (2013) MicroRNA-98 induces an Alzheimer's disease-like disturbance by targeting insulin-like growth factor 1. Neurosci Bull 29(6):745–751
124. Smith P, Al Hashimi A, Girard J, Delay C, Hebert SS (2011) In vivo regulation of amyloid precursor protein neuronal splicing by microRNAs. J Neurochem 116(2):240–247
125. Schonrock N, Ke YD, Humphreys D, Staufenbiel M, Ittner LM, Preiss T et al (2010) Neuronal microRNA deregulation in response to Alzheimer's disease amyloid-beta. PLoS One 5(6):e11070
126. Yan R, Vassar R (2014) Targeting the beta secretase BACE1 for Alzheimer's disease therapy. Lancet Neurol 13(3):319–329
127. Hebert SS, Horre K, Nicolai L, Papadopoulou AS, Mandemakers W, Silahtaroglu AN et al (2008) Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression. Proc Natl Acad Sci U S A 105(17):6415–6420
128. Wang WX, Rajeev BW, Stromberg AJ, Ren N, Tang G, Huang Q et al (2008) The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1. J Neurosci 28(5):1213–1223
129. Wang X, Liu P, Zhu H, Xu Y, Ma C, Dai X et al (2009) miR-34a, a microRNA up-regulated in a double transgenic mouse model of Alzheimer's disease, inhibits bcl2 translation. Brain Res Bull 80(4–5):268–273
130. Alexandrov PN, Dua P, Lukiw WJ (2014) Up-regulation of miRNA-146a in progressive, age-related inflammatory neurodegenerative disorders of the human CNS. Front Neurol 5:181
131. Saba R, Gushue S, Huzarewich RL, Manguiat K, Medina S, Robertson C et al (2012) MicroRNA 146a (miR-146a) is over-expressed during prion disease and modulates the innate immune response and the microglial activation state. PLoS One 7(2):e30832
132. Williams AE, Perry MM, Moschos SA, Larner-Svensson HM, Lindsay MA (2008) Role of miRNA-146a in the regulation of the innate immune response and cancer. Biochem Soc Trans 36(Pt 6):1211–1215
133. Tan L, Yu JT, Liu QY, Tan MS, Zhang W, Hu N et al (2014) Circulating miR-125b as a biomarker of Alzheimer's disease. J Neurol Sci 336(1–2):52–56
134. Liu CG, Wang JL, Li L, Wang PC (2014) MicroRNA-384 regulates both amyloid precursor protein and beta-secretase expression and is a potential biomarker for Alzheimer's disease. Int J Mol Med 34(1):160–166
135. Zi Y, Yin Z, Xiao W, Liu X, Gao Z, Jiao L et al (2014) Circulating microRNA as potential source for neurodegenerative diseases biomarkers. Mol Neurobiol DOI:10.1007/s12035-014-8944-x
136. Ma X, Zhou J, Zhong Y, Jiang L, Mu P, Li Y et al (2014) Expression, regulation and function of microRNAs in multiple sclerosis. Int J Med Sci 11(8):810–818
137. Toivonen JM, Manzano R, Olivan S, Zaragoza P, Garcia-Redondo A, Osta R (2014) MicroRNA-206: a potential circulating biomarker candidate for amyotrophic lateral sclerosis. PLoS One 9(2):e89065
138. Bronze-da-Rocha E (2014) MicroRNAs expression profiles in cardiovascular diseases. Biomed Res Int 2014:985408
139. Thum T, Galuppo P, Wolf C, Fiedler J, Kneitz S, van Laake LW et al (2007) MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure. Circulation 116(3):258–267

140. Chen JF, Murchison EP, Tang R, Callis TE, Tatsuguchi M, Deng Z et al (2008) Targeted deletion of Dicer in the heart leads to dilated cardiomyopathy and heart failure. Proc Natl Acad Sci U S A 105(6):2111–2116
141. Bostjancic E, Zidar N, Stajer D, Glavac D (2010) MicroRNAs miR-1, miR-133a, miR-133b and miR-208 are dysregulated in human myocardial infarction. Cardiology 115(3):163–169
142. Sayed D, Hong C, Chen IY, Lypowy J, Abdellatif M (2007) MicroRNAs play an essential role in the development of cardiac hypertrophy. Circ Res 100(3):416–424
143. Montgomery RL, Hullinger TG, Semus HM, Dickinson BA, Seto AG, Lynch JM et al (2011) Therapeutic inhibition of miR-208a improves cardiac function and survival during heart failure. Circulation 124(14):1537–1547
144. Bonauer A, Carmona G, Iwasaki M, Mione M, Koyanagi M, Fischer A et al (2009) MicroRNA-92a controls angiogenesis and functional recovery of ischemic tissues in mice. Science 324(5935):1710–1713
145. Weber M, Baker MB, Moore JP, Searles CD (2010) MiR-21 is induced in endothelial cells by shear stress and modulates apoptosis and eNOS activity. Biochem Biophys Res Commun 393(4):643–648
146. Nazari-Jahantigh M, Wei Y, Noels H, Akhtar S, Zhou Z, Koenen RR et al (2012) MicroRNA-155 promotes atherosclerosis by repressing Bcl6 in macrophages. J Clin Invest 122(11):4190–4202
147. Wei Y, Nazari-Jahantigh M, Neth P, Weber C, Schober A (2013) MicroRNA-126, −145, and −155: a therapeutic triad in atherosclerosis? Arterioscler Thromb Vasc Biol 33(3):449–454
148. Krutzfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M et al (2005) Silencing of microRNAs in vivo with ‘antagomirs’. Nature 438(7068):685–689
149. Rayner KJ, Suarez Y, Davalos A, Parathath S, Fitzgerald ML, Tamehiro N et al (2010) MiR-33 contributes to the regulation of cholesterol homeostasis. Science 328(5985):1570–1573
150. Rotllan N, Ramirez CM, Aryal B, Esau CC, Fernandez-Hernando C (2013) Therapeutic silencing of microRNA-33 inhibits the progression of atherosclerosis in Ldlr–/– mice--brief report. Arterioscler Thromb Vasc Biol 33(8):1973–1977
151. van Rooij E, Sutherland LB, Thatcher JE, DiMaio JM, Naseem RH, Marshall WS et al (2008) Dysregulation of microRNAs after myocardial infarction reveals a role of miR-29 in cardiac fibrosis. Proc Natl Acad Sci U S A 105(35):13027–13032
152. Soci UP, Fernandes T, Hashimoto NY, Mota GF, Amadeu MA, Rosa KT et al (2011) MicroRNAs 29 are involved in the improvement of ventricular compliance promoted by aerobic exercise training in rats. Physiol Genomics 43(11):665–673
153. Hullinger TG, Montgomery RL, Seto AG, Dickinson BA, Semus HM, Lynch JM et al (2012) Inhibition of miR-15 protects against cardiac ischemic injury. Circ Res 110(1):71–81
154. Li S, Zhu J, Zhang W, Chen Y, Zhang K, Popescu LM et al (2011) Signature microRNA expression profile of essential hypertension and its novel link to human cytomegalovirus infection. Circulation 124(2):175–184
155. Courboulin A, Paulin R, Giguere NJ, Saksouk N, Perreault T, Meloche J et al (2011) Role for miR-204 in human pulmonary arterial hypertension. J Exp Med 208(3):535–548
156. Yang S, Banerjee S, Freitas A, Cui H, Xie N, Abraham E et al (2012) miR-21 regulates chronic hypoxia-induced pulmonary vascular remodeling. Am J Physiol Lung Cell Mol Physiol 302(6):L521–L529
157. Bostjancic E, Glavac D (2014) miRNome in myocardial infarction: future directions and perspective. World J Cardiol 6(9):939–958
158. Corsten MF, Dennert R, Jochems S, Kuznetsova T, Devaux Y, Hofstra L et al (2010) Circulating microRNA-208b and MicroRNA-499 reflect myocardial damage in cardiovascular disease. Circ Cardiovasc Genet 3(6):499–506
159. Kuwabara Y, Ono K, Horie T, Nishi H, Nagao K, Kinoshita M et al (2011) Increased microRNA-1 and microRNA-133a levels in serum of patients with cardiovascular disease indicate myocardial damage. Circ Cardiovasc Genet 4(4):446–454
160. Long G, Wang F, Duan Q, Yang S, Chen F, Gong W et al (2012) Circulating miR-30a, miR-195 and let-7b associated with acute myocardial infarction. PLoS One 7(12):e50926

161. Harvey SJ, Jarad G, Cunningham J, Goldberg S, Schermer B, Harfe BD et al (2008) Podocyte-specific deletion of dicer alters cytoskeletal dynamics and causes glomerular disease. J Am Soc Nephrol 19(11):2150–2158
162. Shi S, Yu L, Chiu C, Sun Y, Chen J, Khitrov G et al (2008) Podocyte-selective deletion of dicer induces proteinuria and glomerulosclerosis. J Am Soc Nephrol 19(11):2159–2169
163. Ho J, Ng KH, Rosen S, Dostal A, Gregory RI, Kreidberg JA (2008) Podocyte-specific loss of functional microRNAs leads to rapid glomerular and tubular injury. J Am Soc Nephrol 19(11):2069–2075
164. Zhdanova O, Srivastava S, Di L, Li Z, Tchelebi L, Dworkin S et al (2011) The inducible deletion of Drosha and microRNAs in mature podocytes results in a collapsing glomerulopathy. Kidney Int 80(7):719–730
165. Sun Y, Koo S, White N, Peralta E, Esau C, Dean NM et al (2004) Development of a microarray to detect human and mouse microRNAs and characterization of expression in human organs. Nucleic Acids Res 32(22):e188
166. Tian Z, Greene AS, Pietrusz JL, Matus IR, Liang M (2008) MicroRNA-target pairs in the rat kidney identified by microRNA microarray, proteomic, and bioinformatic analysis. Genome Res 18(3):404–411
167. Dalla Vestra M, Arboit M, Bruseghin M, Fioretto P (2009) The kidney in type 2 diabetes: focus on renal structure. Endocrinol Nutr 56(Suppl 4):18–20
168. Kato M, Zhang J, Wang M, Lanting L, Yuan H, Rossi JJ et al (2007) MicroRNA-192 in diabetic kidney glomeruli and its function in TGF-beta-induced collagen expression via inhibition of E-box repressors. Proc Natl Acad Sci U S A 104(9):3432–3437
169. Wang Q, Wang Y, Minto AW, Wang J, Shi Q, Li X et al (2008) MicroRNA-377 is up-regulated and can lead to increased fibronectin production in diabetic nephropathy. FASEB J 22(12):4126–4135
170. Zhang Z, Peng H, Chen J, Chen X, Han F, Xu X et al (2009) MicroRNA-21 protects from mesangial cell proliferation induced by diabetic nephropathy in db/db mice. FEBS Lett 583(12):2009–2014
171. Stitt-Cavanagh E, MacLeod L, Kennedy C (2009) The podocyte in diabetic kidney disease. ScientificWorldJournal 9:1127–1139
172. Wang G, Kwan BC, Lai FM, Choi PC, Chow KM, Li PK et al (2010) Intrarenal expression of miRNAs in patients with hypertensive nephrosclerosis. Am J Hypertens 23(1):78–84
173. Zhang W, Zhang C, Chen H, Li L, Tu Y, Liu C et al (2014) Evaluation of microRNAs miR-196a, miR-30a-5P, and miR-490 as biomarkers of disease activity among patients with FSGS. Clin J Am Soc Nephrol 9(9):1545–1552
174. Denby L, Ramdas V, McBride MW, Wang J, Robinson H, McClure J et al (2011) miR-21 and miR-214 are consistently modulated during renal injury in rodent models. Am J Pathol 179(2):661–672
175. Ichii O, Otsuka S, Sasaki N, Namiki Y, Hashimoto Y, Kon Y (2012) Altered expression of microRNA miR-146a correlates with the development of chronic renal inflammation. Kidney Int 81(3):280–292
176. Lu J, Kwan BC, Lai FM, Tam LS, Li EK, Chow KM et al (2012) Glomerular and tubulointerstitial miR-638, miR-198 and miR-146a expression in lupus nephritis. Nephrology (Carlton) 17(4):346–351
177. Wang G, Kwan BC, Lai FM, Choi PC, Chow KM, Li PK et al (2010) Intrarenal expression of microRNAs in patients with IgA nephropathy. Lab Invest 90(1):98–103
178. Sui W, Yang M, Li F, Chen H, Chen J, Ou M et al (2014) Serum microRNAs as new diagnostic biomarkers for pre- and post-kidney transplantation. Transplant Proc 46(10):3358–3362
179. Deltas C, Papagregoriou G (2010) Cystic diseases of the kidney: molecular biology and genetics. Arch Pathol Lab Med 134(4):569–582
180. Lee SO, Masyuk T, Splinter P, Banales JM, Masyuk A, Stroope A et al (2008) MicroRNA15a modulates expression of the cell-cycle regulator Cdc25A and affects hepatic cystogenesis in a rat model of polycystic kidney disease. J Clin Invest 118(11):3714–3724

181. Pandey P, Brors B, Srivastava PK, Bott A, Boehn SN, Groene HJ et al (2008) Microarray-based approach identifies microRNAs and their target functional patterns in polycystic kidney disease. BMC Genomics 9:624
182. Patel V, Williams D, Hajarnis S, Hunter R, Pontoglio M, Somlo S et al (2013) miR-17~92 miRNA cluster promotes kidney cyst growth in polycystic kidney disease. Proc Natl Acad Sci U S A 110(26):10765–10770
183. Agostini M, Knight RA (2014) miR-34: from bench to bedside. Oncotarget 5(4):872–881
184. Zhao X, Pan F, Holt CM, Lewis AL, Lu JR (2009) Controlled delivery of antisense oligonucleotides: a brief review of current strategies. Expert Opin Drug Deliv 6(7):673–686
185. Shimakami T, Yamane D, Welsch C, Hensley L, Jangra RK, Lemon SM (2012) Base pairing between hepatitis C virus RNA and microRNA 122 3′ of its seed sequence is essential for genome stabilization and production of infectious virus. J Virol 86(13):7372–7383
186. Lindow M, Kauppinen S (2012) Discovering the first microRNA-targeted drug. J Cell Biol 199(3):407–412
187. Elmen J, Lindow M, Schutz S, Lawrence M, Petri A, Obad S et al (2008) LNA-mediated microRNA silencing in non-human primates. Nature 452(7189):896–899
188. Lanford RE, Hildebrandt-Eriksen ES, Petri A, Persson R, Lindow M, Munk ME et al (2010) Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. Science 327(5962):198–201
189. Janssen HL, Reesink HW, Lawitz EJ, Zeuzem S, Rodriguez-Torres M, Patel K et al (2013) Treatment of HCV infection by targeting microRNA. N Engl J Med 368(18):1685–1694
190. Liu C, Kelnar K, Liu B, Chen X, Calhoun-Davis T, Li H et al (2011) The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. Nat Med 17(2):211–215
191. Craig VJ, Tzankov A, Flori M, Schmid CA, Bader AG, Muller A (2012) Systemic microRNA-34a delivery induces apoptosis and abrogates growth of diffuse large B-cell lymphoma in vivo. Leukemia 26(11):2421–2424
192. Wiggins JF, Ruffino L, Kelnar K, Omotola M, Patrawala L, Brown D et al (2010) Development of a lung cancer therapeutic based on the tumor suppressor microRNA-34. Cancer Res 70(14):5923–5930
193. Bader AG (2012) miR-34 – a microRNA replacement therapy is headed to the clinic. Front Genet 3:120

Chapter 3
piRNAs-Transposon Silencing and Germ Line Development

Catherine Demoliou

Introduction

DNA in eukaryotic cells is packaged into two forms of chromatin: euchromatin, and heterochromatin. Euchromatin is rich in actively transcribed genes. Heterochromatin is rich instead, in repeated, non-coding sequences representing multiple copies of ‘selfish’ or “parasitic” intergenic genetic elements that are epigenetically repressed [1–3]. These elements, called transposons or transposable elements (TEs), are able to mobilize “jump” within the genome, and to multiply during their transposition. TEs represent 45 % of the human and primate genomes. They are mainly scattered between and within genes in mammals including humans, and at constitutive heterochromatin pericentromeric and subtelomeric regions in plants and in *Drosophila* fruit flies. There are two major TE classes. Class I includes the autonomous retrotransposons, which either have long interspersed nuclear elements (LINEs), or no long terminal repeats (non-LTRs). Both types replicate in a “copy-and-paste” manner. That is, a TE is transcribed first to messenger RNA (mRNA), which associates with self-encoding, reverse transcriptases and endonucleases in the cytoplasm, and then it is transported back into the nucleus. There, the TE-RNA is reverse transcribed to DNA and integrated into a new site in the host genome. Subtypes of LTRs, the non-autonomous LTRs, contain short interspersed elements (SINEs), which replicate using the activities of an endonuclease and of a reverse transcriptase encoded either within the LINE-1 elements (L1), or within other TEs. In contrast, the Class II TEs are of the DNA-type. They transpose via a DNA “cut-and-paste” mechanism using the “target capture” action of a transposase, which is often encoded within the DNA sequence of the TE.

C. Demoliou, Ph.D. (✉)
Life and Health Sciences Department, School of Sciences and Engineering,
University of Nicosia, 46 Makedonitissas Ave., PO Box 24005, 1700 Nicosia, Cyprus
e-mail: demoliou.c@unic.ac.cy

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_3

TEs have contributed to genomic changes that had an impact on the evolution of eukaryotes, and are responsible for intra-population genetic variations. Recent evidence shows that through evolution, TEs, once thought to be only parasites, have been recruited by genomes to defend TE invasions. TE transpositions, however, continue to be a thread, since TEs have not stopped replicating and mobilizing independently within the genome. Mutations caused by TE transpositions may result in the deregulation of transcription and translation of genes required for normal cell function. Furthermore, TE transpositions can be the cause for diseases including cancer. *De novo* transpositions of TEs are linked to at least 65 known human diseases. Most importantly, TE transpositions during germ line development can be the cause of developmental, hereditary and fertility disorders [1–6].

The control of gene expression in eukaryotic organisms is by gene silencing at DNA level, involving chromatin remodelling that is regulated via DNA methylation and epigenetic modifications [4–6]. In addition, eukaryotes have evolved highly efficient and specialized systems, which detect, guide, and repress the expression of genes and especially of active TEs. Such systems recruit non-coding RNAs to regulate TE-expression via transcriptional gene silencing (TGS), or posttranscriptional gene silencing (PTGS). TGS systems operate in the nucleus and can induce epigenetic modifications guided by long or short antisense non-coding RNAs. PTGS systems inactivate or degrade targeted mRNA or TEs. They operate mainly in the cytoplasm where they are guided by short non-coding RNAs. These small RNAs associate with effector proteins to form the catalytic core of the RNA-induced silencing complex (RISC) [7–9]. In addition, as we will see later on, there are other partner proteins, which contribute to the assembly of a functional RISC. The degradation of targeted mRNA via RISC, is called RNA interference (RNAi), and it was first observed in plants [10, 11].

The small RNAs used for TGS or PTGS in metazoan include several classes: the micro RNAs (miRNAs), which regulate gene expression; the small interfering RNAs (siRNAs), which regulate gene expression and TE transpositions, and the small non-coding P-element induced RNAs (piRNAs), which repress TE expression and mobilization specifically in germ line cells during gametogenesis (Fig. 3.1). piRNAs (24–32 nt long) exert their effect by binding to the P-element induced wimpy testis (PIWI) subfamily of proteins of the Argonaute family. The smaller miRNAs and siRNAs (20–23 nt long), exert their effects by binding instead to the AGO subfamily members of the Argonaute family [14–17].

This chapter will concentrate on the biological roles of piRNAs and the PIWI proteins, which are expressed in germ cells and form the catalytic core of piRISC. Genetic studies on PIWI proteins using animal models and the isolation and sequencing of piRNAs have aided in building up a coherent picture of piRISC functions during gametogenesis [12, 18–25]. This picture, although not complete, shows the evolution of interrelationships between piRNAs and TEs expressed during embryogenesis and postnatal germ cell development. Germ cells use TEs in *cis*, as the source of small RNA regulatory sequences (e.g. piRNAs) to guide the degradation of TEs or mRNA. TEs in *trans*, appear to provide instead RNA regulatory sequences for modulating gene transcription/translation (e.g. DNA methylation, RNA inactivation by deadenylation etc.) [1–3, 22–25].

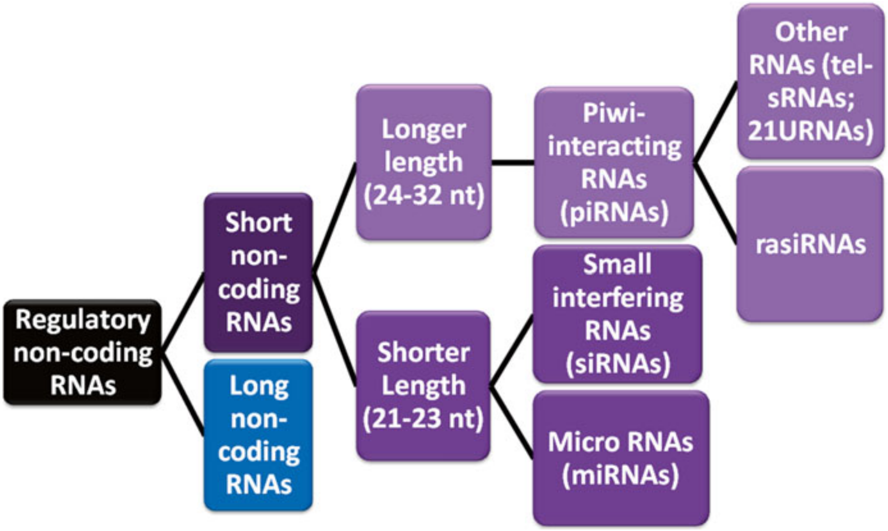

Regulatory non-coding RNAs
Short non-coding RNAs
Long non-coding RNAs
Longer length (24-32 nt)
Shorter Length (21-23 nt)
Piwi-interacting RNAs (piRNAs)
Small interfering RNAs (siRNAs)
Micro RNAs (miRNAs)
Other RNAs (tel-sRNAs; 21URNAs)
rasiRNAs

Fig. 3.1 Classes of regulatory non-coding RNAs and subsets of piRNAs. Long non-coding RNAs refer to the antisense RNAs involved in epigenetic functions [7, 8]. piRNA subsets: repeat-associated-siRNAs (rasiRNAs) identified in organisms like *Drosophila* and Zebrafish [12]; mouse heterochromatin associated pi-like small RNAs (tel-sRNAs) [13]; 21U RNAs, small RNAs expressed in the *C. Elegans* germ line [14]

Germ Cells and Embryonic Development

Mammalian germ cells arise relatively late during embryo development from a small population of extra-embryonic mesoderm cells of the epiblast. The cells develop into primordial germ cells (PGCs) that migrate and colonize the genital ridges that will form the gonads. Sex specification of PGCs in foetal testes, is defined by paternally inherited determinants expressed in the early development of the zygote. In the ovaries, it is defined by maternally inherited determinants present in the germ plasma of the oocytes. During ovary development, the PGCs first proliferate via mitosis, and then differentiate into germline stem cells (GSCs), known as secondary oocytes (Fig. 3.2). Subsequently, GSCs enter meiosis I and arrest in prophase. The completion of meiosis I occurs only upon onset of sexual maturation and ovulation when again they arrest in metaphase of meiosis II. At this stage, the oocytes build-up special cytoplasmic granules, the so-called germ plasma that is required to support the embryo for the next generation of PGCs. In addition, the presence of GSCs drives the development of support somatic (follicle) cells. In mice, the mature oocytes remain transcription-silent until fertilization occurs. Fertilization signals the oocyte meiosis to resume, and the highly differentiated oocytes finally transform into totipotent embryos [26].

Male mammalian PGCs, in contrast, undergo only mitotic proliferation during the stages of embryonic development (Fig. 3.2). They then arrest in G1 phase (spermatogonia), and differentiate to GSCs in the male gonad. Mitotic proliferation and

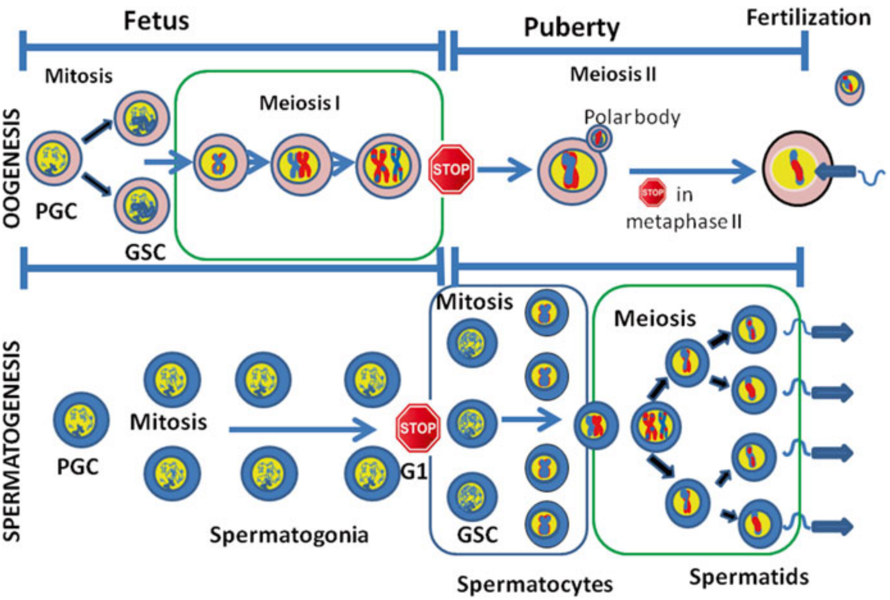


Fig. 3.2 Developmental germ cell transitions during gametogenesis. A schematic diagram showing mammalian female and male gametogenesis in the fetus, and adulthood [26]. In ovaries the number of primary oocytes is determined during foetal development. There is no further proliferation after birth; meiosis starts with sexual maturation, and is completed only upon fertilization. Male germ cells continue to proliferate upon sexual maturation to form spermatocytes that undergo meiosis to form haploid spermatids; *PGS* primordial germ cell, *GSC* germ stem cells

differentiation of spermatogonia into spermatocytes resume in the testes, only upon sexual maturation. Adult spermatocytes undergo meiosis to form each, four haploid spermatids that lose their cytoplasm as they develop into mature spermatozoa (sperm) able to fertilize an oocyte [26].

Mitotic proliferations of PGCs and transition into GSCs, prenatally in females, and during juvenile development in males, are characterized by a genome-wide DNA CpG demethylation by DNA methyltransferase (DNMT) enzyme(s). These processes require the expression of pluripotent stem cell markers (i.e. Oct4, Vasa, Fragilis and Nanog) [27], as well as several developmental factors, which define whether the cells will differentiate (specification) or continue to proliferate as pluripotent GSCs [28]. DNA demethylation is required in order to erase inherited imprints from both parents, and to ensure the pluripotency of the GSCs. It is also required to enable genomic reprogramming and resetting of imprinting for germ cell specification [6, 26]. It is this period of DNA demethylation that TEs can be particularly active due to the relaxation of epigenetic control and to the increased expression of transcription factors [2].

The maintenance of the GSC phenotype and GSCs proliferation involve extensive changes in histone modifications and RNA-driven gene silencing processes, which are specific to the germ line. Signals and molecules that commit germ cells to undergo meiosis are provided by somatic support cells. Meiotic progression events (i.e. crossing over, meiotic silencing of unpaired chromatin, imprinting), are also guided by epigenetic transitions that involve the re-establishment of DNA methylation patterns (*de novo* DNA methylation) as well as the post-translational modification of major histones [27–30].

The RNA-driven gene silencing processes are closely associated with the maternally inherited piRNAs and with the PIWI-dependent biogenesis of regulatory piRNAs from RNA-precursor transcripts [21]. During the window period of DNA demethylation and during germ cell mitotic self-renewal and meiosis, the piRNAs (inherited and newly generated), bind PIWI to form piRISC for targeted TGS or PTGS [21, 31–37]. These processes require the support of Tudor-domain-containing proteins known as Tudor domain-related proteins (TDRDs), which form the cytoplasmic scaffold for piRNA loading and processing and for piRISC mobilization and function [38–49]. In addition, a functional piRNA-PIWI pathway is also required for ovarian somatic cell-support during germ cell development in the zygote [34, 50]. The PIWI protein expression and the piRNAs signatures identified in diverse species such as *Drosophila*, zebra fish and mammals, indicate that this pathway has been conserved through evolution [51].

Other short RNAi pathways, i.e. the miRNA pathway and the siRNAs-pathway, are also important for the proper development of germ cells. Recent findings suggest that PTGS during the very early developmental stages of the embryo, between oocyte fertilization and blastocyst implantation, may involve RNAi processes that are guided by temporal transitions of endogenous small RNAs, from retrotransposon-derived siRNAs/piRNAs or zygote synthesized miRNAs [52, 53] (Fig. 3.3).

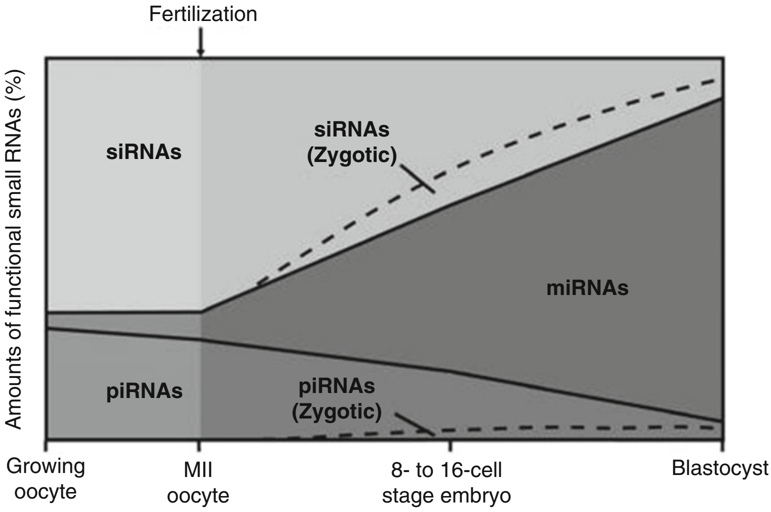


Fig. 3.3 Proposed transition of functional small RNAs during mammalian oogenesis and early embryogenesis. The schematic diagram is not drawn to scale. *Dotted lines* represent the putative expression of zygotic siRNAs and piRNAs. Reprinted from Ohnishi Y, et al. Small RNA class transition from siRNA/piRNA to miRNA during pre-implantation mouse development. Nucleic Acids Res. 2010; 38(15):5141–5151 (Open Access) [52]

The PIWI Proteins

Most of the information about the function of PIWI proteins comes primarily from studies in *Drosophila* and mice. The *Drosophila* PIWI proteins Aubergine (Aub), Piwi and Argonaute-3 (Ago3) comprise the PIWI subfamily of the highly conserved Argonaute family. The other subfamily, referred to as AGO, includes the Ago1 and Ago2 proteins that bind miRNA and siRNA, respectively [40, 54–56]. AGO proteins are ubiquitously expressed and have been found in almost all eukaryotes. Metazoans have representatives of both Argonaute subfamilies. Fungi, green algae and plants encode exclusively AGO-like proteins whereas amoebas encode exclusively PIWI-like proteins, placing *piwi* as the oldest ancestor gene and first in the line of evolution of the Argonaute clan [12, 14–17, 46, 51].

In *Drosophila*, the Ago3 and Aub proteins are expressed primarily in the cytoplasm of male and female germ cells and in the cytoplasm of female somatic nurse cells. In contrast, Piwi is a nuclear protein and it is expressed in both, germline and ovarian somatic cells [54]. PIWI proteins are required for the biogenesis of *Drosophila* piRNAs and for the piRISC-mediated degradation of TEs in germ cells, as well as for TGS in somatic and in germ cells [21]. These activities are interlinked and essential for germ line maintenance, proliferation and differentiation that are required to ensure fertility. Piwi and Aub expression is required for ensuring both male and female fertility whereas Ago3 expressions seems to be more important in female fertility [40, 57]. Mutations in PIWI proteins result in the over expression and mobilization of retrotransposons in male flies. This results in DNA damage and germ cell apoptosis that causes sterility since GSC maintenance, proliferation and embryonic axis specifications are defective [20–22, 24, 25, 32–34, 50, 54–56].

The Mouse PIWI Family

The mouse Miwi (*Piwil1*), Mili (*Piwil2*) and Miwi2 (*Piwil4*) proteins, share significant homology with their *Drosophila* and human counterparts. Their expression in different, overlapping periods during gametogenesis, is specific to the male germ line (Fig. 3.4). Mili is expressed at E12.5 of embryonic stage, and is present throughout gametogenesis to the round post-meiotic spermatid stage; Miwi2 is transiently expressed at E15.5 of embryonic stage until shortly after birth [41], and Miwi is expressed from the pachytene meiotic stage to post-meiotic round spermatid stage in the adult testis [42, 55, 58]. Mili binds piRNAs of 24–28 nt long, whereas Miwi2 and Miwi bind piRNAs that are a little longer (27–32 nt) [36, 58]. Studies have shown that Mili plays a role in TE control, GSC maintenance and differentiation; Miwi2 in TE control and genocyte proliferation, and Miwi in meiosis during spermatogenesis [20–23]. Their differential role has also been suggested from evidence that shows binding (Miwi2, Mili) to prenatal piRNAs with nucleotide sequences complementary to TEs and binding (Miwi, Mili) to TE-derived postnatal piRNA or to non-coding intergenic and genic piRNAs during spermatogenesis (pre and post meiosis) [35, 47, 48].

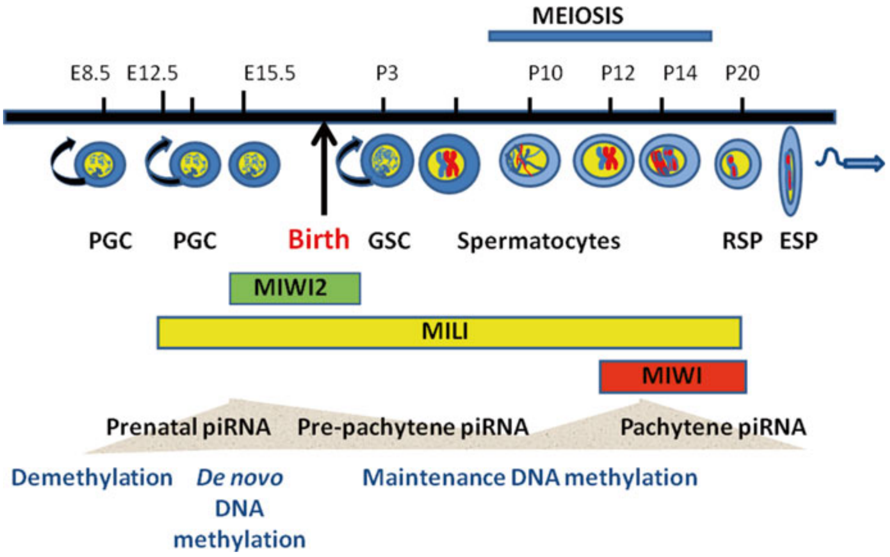


Fig. 3.4 PIWI protein and piRNA expression during mammalian spermatogenesis. A schematic representation of mouse spermatogenesis on a time coordinate showing the periods during which Mili, Miwi2, Miwi and piRNA are expressed and epigenetic reprogramming takes place (i.e. DNA demethylation/methylation) [2]. *PGC* primordial germ cells, *GSC* germ stem cells, *RSP* round spermatids, *ESP* elongated spermatids. *Bent arrows* indicate cell self-renewal

Several studies using PIWI mutant/deficient mice have provided an inside of PIWI protein functions. Prenatal Miwi2 and Mili expressions have been associated with piRNA-guided *de novo* DNA methylation that represses TE expression as well as with the piRNA-targeted cleavage of TE-RNA [20–23, 58–62]. If Miwi2 or Mili proteins are not expressed in prenatal germ cells, the genomic regions of Line1 and intracisternal A-particle (IAP) type transposons (non-LTR) are hypomethylated. As a result, TE mRNA levels are up-regulated causing spermatogenic stem cell arrest at leptotene (Miwi2) or zygotene/early pachytene (Mili) stages that results in germ cell apoptosis and sterility [45, 47, 48, 58]. Since Mili and Miwi2-deficient mice have impaired prenatal piRNA production, Mili and Miwi2 are considered to be involved in piRNA biogenesis [63]. Deficiency in Miwi, which also results in sterility, affects adult germ cells instead, by arresting spermatogenesis and thus preventing spermiogenesis [35, 64]. Furthermore, ribonucleoprotein-associated mRNA that is required for GSC renewal and for spermiogenesis are stabilized by Mili and Miwi, respectively [59, 63, 65].

In contrast to the male germ line, female mouse germ cells express only Mili. Studies with female knockout mice for Mili and/or proteins that associate with piRISC, have shown that female fertility is not affected. These sex-specific differences in mice suggest that the expression of PIWI members and/or associated proteins may be redundant for mammalian oogenesis. Mammalian oogonia, unlike germ cells, proliferate and undergo meiosis only during gestation (Fig. 3.2), and they may have an adequate amount of maternally inherited piRNAs for controlling TE expression [21, 35, 41, 45]. Alternatively, transposon control in mammalian oocytes, may involve other non-coding RNAs, like siRNAs, or proteins that block TE integration and/or provide innate immunity against retroviruses [52, 53, 66–68].

The specific requirements for PIWI function and the mechanisms of TE silencing observed in the male germ line may be unique to mammals and the result of the evolution of the mammalian germ cell development system. In other vertebrates (i.e. *Drosophila* flies, aphids and zebra fish), and in distant organisms (i.e. sea urchin), both, male and female gametogenesis require the contributions of all PIWI members for normal cell specification, maintenance and germ cell meiosis [69]. In *Drosophila*, for example, maternally inherited piRNAs may contribute to female PGCs formation and GSCs specification during early embryogenesis. However, unlike mammals, GSC self-renewal divisions in *Drosophila*, requires a functional Piwi that resides in the nurse and somatic cells of the ovary [34, 50, 55, 70, 71].

The Human PIWI Family

All known members of the AGO family have been identified in the human genome [72]. PIWI proteins are expressed mainly in human testes and consist of the Piwil1/Hiwi, Piwil4/Hiwi2, Piwil3/Hiwi3 and the Piwil2/Hili proteins with homologues in other mammals and vertebrates. The *PIWL1, PIWIL4, PIWIL3* and *PIWIL2* genes are on chromosomes 12, 11, 22 and 8, respectively. The three AGO genes (*AGO1, AGO3, and AGO4*) are closely clustered on chromosome 1 suggesting a common evolution from concurrent gene duplications originating from the *AGO2* gene (of a more ancestral origin) found on chromosome 8 [73]. In normal human testes, Hiwi (the Miwi protein homologue), is specifically expressed in spermatocytes and round spermatids during spermatogenesis, and its over-expression is associated with seminoma tumours [74]. Hiwi2/Miwi2, the only protein member expressed in human somatic cells ubiquitously, may play a role in chromatin remodelling [75].

The Endonuclease Activity of the PIWI Proteins

The members of the PIWI subfamily, like those of the AGO subfamily, are characterized by the conserved PAZ (Piwi-Argonaute-Zwille) domain, the PIWI sequence domains and a middle (MID) sequence domain (Fig. 3.5). The PAZ and PIWI-MID domains are responsible for facilitating the formation of the double stranded RNA complex (i.e. small RNA bound to single stranded target mRNA), required for the cleavage of target RNA. The MID-domain recognizes and binds specifically the characteristic 5′ end phosphate of Uridine (1U-bias) of piRNA-precursor molecules or piRNAs generated products, and ensures the correct orientation of the bound piRNA by anchoring it. The PAZ domain at the N-terminal region, forms a pocket that binds the 3′ overhang of the piRNA. Cleavage of mRNA occurs at the PIWI

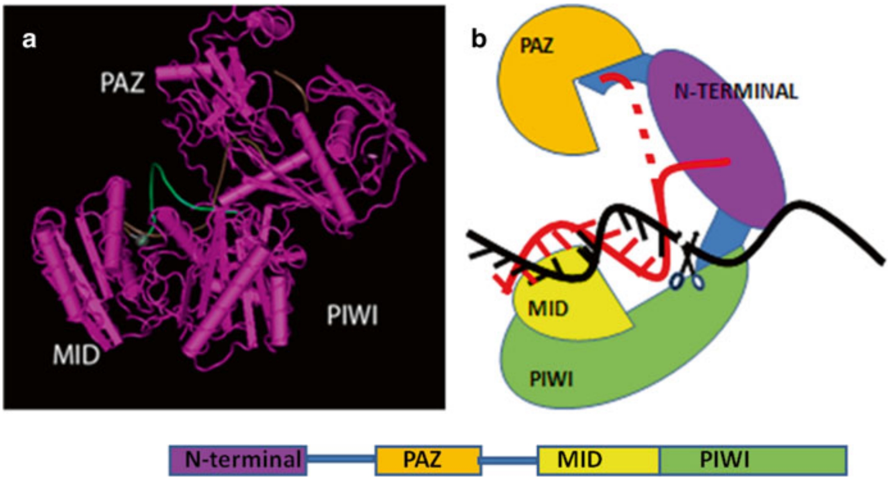


Fig. 3.5 Protein structure and a putative model of PIWI slicer activity based on Ago structure. (**a**) The crystal structure of *Thermus thermophilus* Ago, showing the PAZ, MID and PIWI domains (PDB identity 3F73; created using Cn3D). (**b**) A putative model of PIWI-bound piRNA (*red*) forming a double strand with target-RNA (*black*) at the catalytic pocket of PIWI. The 5′ end of piRNA is stabilized by the MID domain. The *solid* and *dotted lines* at the 3′ end of piRNA denote a PIWI conformational change after base pairing that is considered to trigger PIWI cleavage of target RNA [76]

domain, which is at the C terminal region and contains the catalytic DDH triad composed of two aspartic acid residues (D) and one histidine (H). This catalytic site has an RNAse H-like "slicer" endonuclease activity responsible for the dsRNA guided hydrolysis of target mRNA bound to piRNA (Fig. 3.5) [76–79].

The endogenous PIWI-slicer activity of some PIWI proteins is probably responsible for both, the biogenesis of regulatory mature piRNAs and the piRNA-targeted TE degradation [35, 37, 63–67, 79]. The first phase of the piRNA-PIWI pathway involves the pre-cleavage of very long RNA transcripts to fragments of various lengths (primary piRNA-precursors), to which PIWI proteins may bind via the pre-bound piRNAs. The intermediates formed are then cleaved into smaller fragments by the PIWI slicer activity, trimmed to size by a 3′ to 5′ exonuclease (i.e. Ago2, which co-localizes with Ago3 and Miwi in the chromatoid body of male germ cells), and subsequently 2′-O-methylated 3′ by the HEN1 methyl transferase to form the mature and stable "primary piRNAs" with regulatory function(s) [20, 32, 63, 65, 75–82]. The 5′ to 3′ processing of primary piRNA-precursors is considered to be undertaken by an AGO subfamily member [20, 81, 82]. However, in vivo studies using the method of cross-linking prior to immunoprecipitation (HITS-CLIP method) of PIWI-piRNA complexes, have suggested instead, that long primary piRNA-precursors are processed to mature piRNAs via a 5′ to 3′ endonuclease activity of PIWI proteins (Mili, Miwi) without being guided by pre-bound piRNA [65]. Furthermore, studies with mouse Miwi-catalytic mutants of the DDH triad and with Miwi deficient mice have indicated that the biogenesis of primary piRNA-precursors and primary piRNAs is independent of PIWI-slicer activity [63].

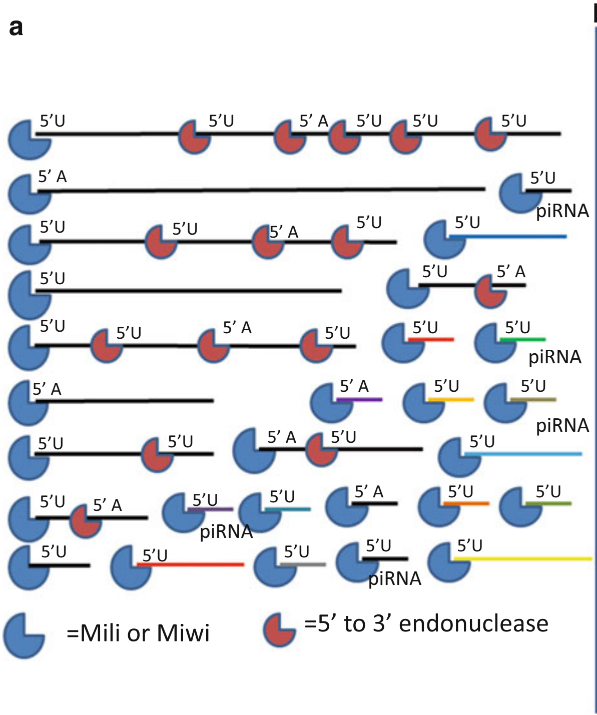


Fig. 3.6 A schematic representation of a putative mechanism for primary biogenesis of piRNAs. (**a**) PIWI (*blue packman*) binds non-coding RNA through the 5′ end U and promotes its 5′ to 3′ cleavage by an endonuclease (*red packman*). The fragments generated continue to be loaded onto PIWI to be further cleaved and/or as they get smaller to be finally trimmed to the right piRNA size that bind PIWI with high affinity and gets 2′-O-methylated at the 3′ end to mature piRNAs with diverse sequences (shortest lines in different colours). (**b**) Changes in the 3D structure of a random sequence RNA fragment as it is cleaved by a 5′ to 3′ endonuclease while its 5′ end is bound to PIWI. *Parallel lines* show base pairing

Based on current evidence an alternative hypothesis could be that Ago2 or another endonuclease may cleave very long mRNA transcripts that are bound with low affinity to PIWI proteins via their 5′U/A ends (Fig. 3.6a). After each round of endonuclease cleavage in a 5′ to 3′ direction, the 5′U end fragments generated will have unique secondary or 3D structures and upon binding to PIWI may now provide different new sites for endonuclease cleavage (Fig. 3.6b). Depending on their size, the PIWI-bound fragments would continue either to be further cleaved by the endonuclease, or trimmed by a 3′ to 5′ exonuclease. Those with the right size to bind PIWI with higher affinity (i.e. by 3′ binding to the PAZ domain) may then form a stable piRNA-PIWI complex, which can be methylated by HEN1 methyltransferase to produce mature piRNAs having primarily U at the 5′-end but diverse sequences (Fig. 3.6a). On the basis of such a model, Mili and Miwi, may participate and speed up primary piRNA biogenesis by stabilizing the 5′ end of single stranded piRNA precursors without using their endogenous slicer activity [65]. This slicer activity, in fact, could be required only for secondary processing as suggested from studies on LINE1 transposon silencing during spermatogenesis [63, 83]. Nuclear Miwi2 is most likely to be involved instead in TGS, by regulating DNA methylation of target transposon loci [58–61]. As mentioned above, the cooperation of several other proteins, especially of TDRDs, is needed for the loading of piRNAs intermediates into Piwi as well as for acting as chaperons (i.e. heat shock protein 90, HSP90, homologues) for piRNAs mobilization and the regulation of TE and/or transcription of precursor piRNA [33, 49, 66, 84–89].

The N-terminal region of PIWI proteins, rich in R-G or R-A amino acids, is the site for post transcriptional methylations by symmetrical arginine methyltransferase (SAM) enzymes that are important in specification and maintenance of germ cells. The symmetrically dimethylated (sDMS) status of the N-terminal of PIWIs (Fig. 3.7), may determine the interactions with components involved either in piRNA biogenesis or in macromolecular assembly required for piRISC mobilization and function [90–92]. The stable association of the mature primary piRNA

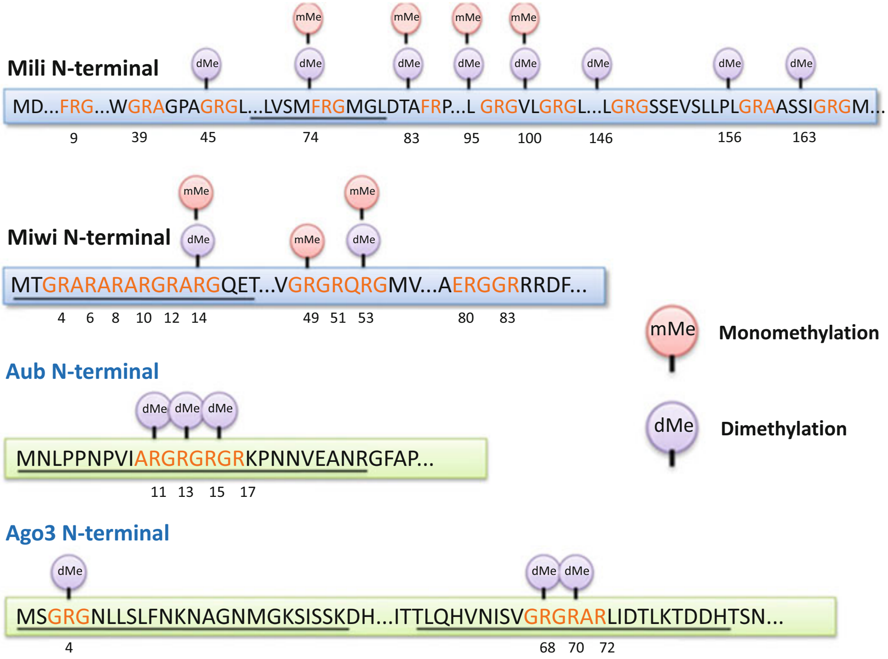
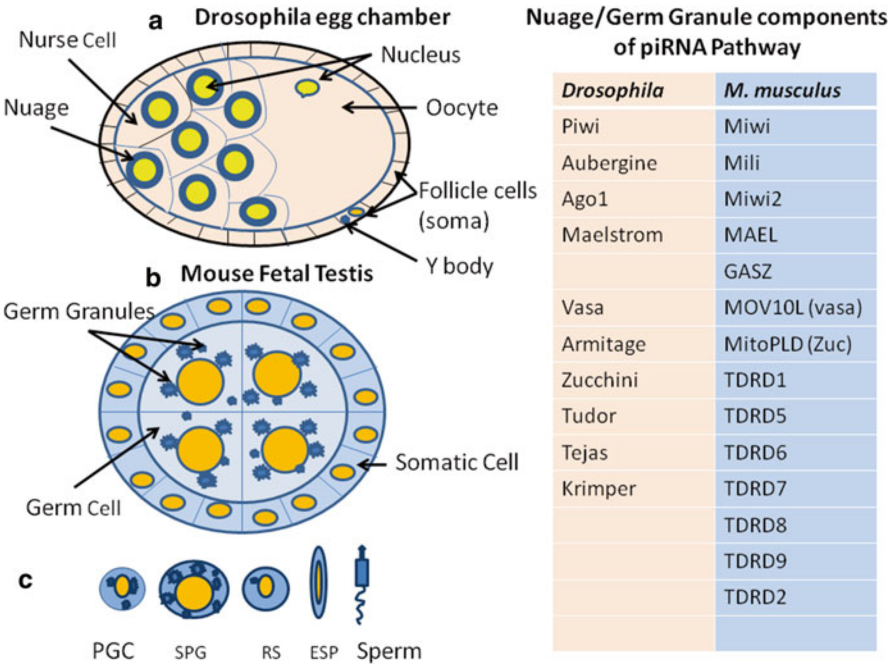

Mili N-terminal
mMe
dMe
MD...FRG...WGRAGPAGRGL...LVSMFRGMGLDTAFRP...L GRGVLGRGL...LGRGSSEVSLLPLGRAASSIGRGM...
9 39 45 74 83 95 100 146 156 163
Miwi N-terminal
MTGRARARARGRARGQET...VGRGRQRGMV...AERGGRRRDF...
4 6 8 10 12 14 49 51 53 80 83
Aub N-terminal
MNLPPNPVIARGRGRGRKPNNVEANRGFAP...
11 13 15 17
Ago3 N-terminal
MSGRGNLLSLFNKNAGNMGKSISSKDH...ITTLQHVNISVGRGRARLIDTLKTDDHTSN...
4 68 70 72
mMe Monomethylation
dMe Dimethylation

Fig. 3.7 Arginine methylation status of PIWI proteins. The N-terminal sequences of mouse PIWI (Mili and Miwi) and fly PIWI (Aub and Ago3) are shown with putative sDMA motifs (in *pink*). Identified methylation sites (dimethylation [dMe] or monomethylation [mMe]) are shown above the relevant arginine, with the residue numbers below. *Underlined* sequences indicate the synthetic peptides used in studies for pull-down assays. Reprinted from Siomi MC, Mannen T, Siomi H. How does the royal family of Tudor rule the PIWI-interacting RNA pathway? Genes Dev. 2010; 24:636–646. Review. (Open Access)

with Piwi may induce a conformation change at the N-terminal of Piwi, which dictates the methylation pattern of PIWI by the symmetric arginine dimethylase enzyme (dPRMT5). This pattern in turn, may define which TDRD protein binds to PIWI to modulate and/or guide piRISC function(s) specified by the signature of the bound piRNA (i.e. PIWI localization, piRNA biogenesis, mobilization to the nucleus for TE silencing, and/or the regulation of gene transcription/translation) [37, 55, 85, 86, 92–103].

During *Drosophila* embryogenesis, piRISC-associate proteins co-localize with PIWI proteins in the oocyte germ plasma and in cell specific cytoplasmic granules (nuage), in nurse and ovarian somatic cells (Fig. 3.8). The precise molecular function of these proteins and the mechanisms that ensure binding specificities are not as yet, known. However, loss of TDRD protein function results in the expressions of TEs, impaired gametogenesis, meiosis arrest and sterility. In mammalian cells, like in *Drosophila*, a larger number of such associate TDRDs, co-localize with PIWI in similar germ and somatic cell granules (Fig. 3.8). Their pattern of expression is highly correlated with PIWI expression, and it appears that distinct combinations



Nuage/Germ Granule components of piRNA Pathway

Drosophila	*M. musculus*
Piwi	Miwi
Aubergine	Mili
Ago1	Miwi2
Maelstrom	MAEL
	GASZ
Vasa	MOV10L (vasa)
Armitage	MitoPLD (Zuc)
Zucchini	TDRD1
Tudor	TDRD5
Tejas	TDRD6
Krimper	TDRD7
	TDRD8
	TDRD9
	TDRD2

Fig. 3.8 Schematic representation of (**a**) *Drosophila* ovarian egg chamber (**b**) mouse fetal testis and (**c**) mouse germ cells. Tudor (TUD), tudor-domain containing proteins (TDRD), developmental factors and PIWI proteins that accumulate to form macromolecular assemblies in germ granules (*blue*) and in the nuage (*blue*) are listed in the Table for *Drosophila* and mouse. Adapted from Chen C, Nott TJ, Jin J, Pawson T. Deciphering arginine methylation: Tudor tells the tale. Nat Rev Mol Cell Biol 2011; 12:629–642 with permission from Nature Publishing Group

of TDRDs-PIWI assemblies characterize transposon silencing and spermiogenesis. Targeted deletions of TDRDs do not affect germ cell viability but they do affect the maintenance of GSCs and male fertility [49, 100–103]. There are also a number of other developmental markers, which co-localize with PIWI proteins in the cytoplasmic granules and are just as important for GSC development and spermatogenesis (Fig. 3.8). The PIWI associations with the various proteins including TDRDs in cytoplasmic granules, appear to be conserved through evolution [21, 90, 91, 94–96, 101].

There is limited knowledge at present as to how the expression of PIWI proteins is regulated during gametogenesis. Hou et al. [104] have shown that the expression of *Miwi* from mid pachytene spermatocyte stage to round spermatid stage, is controlled by developmental transcription factors through a DNA methylation-dependent mechanism involving epigenetic modifications at the putative PIWI promoter region. Regulation of PIWI protein expression by developmental transcription factors would be in agreement with the fact that PIWI proteins are highly expressed only in germ line cells during gametogenesis.

PIWI Evolution

The RNAi system of defence against retrotransposons is very ancient. A plausible direct evolutionary connection between prokaryotes and the RNAi system in eukaryotes, is supported by the evidence for the existence of a prokaryotic immune system that uses as guides RNA or DNA molecules and homologues of the AGO family to degrade nucleic acids of invading elements [105]. Furthermore, organisms from far back in evolution, such as sponges, and organisms near the evolutionary basis of metazoans, express *piwi*-like genes in somatic stem cells that give rise to the next generation, in support of a PIWI ancestral origin and its ancient role in gametogenesis [106–110].

The diversification of the AGO gene family into the AGO and PIWI families in contemporary animals including humans, is considered to have been established since the origin of metazoan, (i.e. more than 600 million years ago), and it must have played an important role in the evolution of multicellularity [15, 73, 105–108]. Phylogenetic analysis shows that after branching out, both, the AGO and the PIWI proteins have undergone a marked degree of expansion. Both must have evolved following vertebrate-specific duplication events, which were independent and lineage-specific (Fig. 3.9), in agreement with the reported absence of pairwise orthologies between the PIWI members of *Drosophila* and mice. The *Drosophila* PIWI proteins (Piwi, Aub and Ago3), for example, are related structurally more to each other rather than to the mouse PIWI family members (Mili, Miwi and Miwi2) [105]. In terms of the evolution of PIWI structure, the PAZ domains of Piwi-like and of Dicer-like proteins appear to cluster together, whereas the PAZ domain of the AGO members appears to have evolved separately. This suggests that the multidomain formation in AGO members may have occurred after an *ago-piwi* gene duplication with AGO inheriting the PAZ domain from an ancestral *piwi* gene [15, 105, 108, 111].

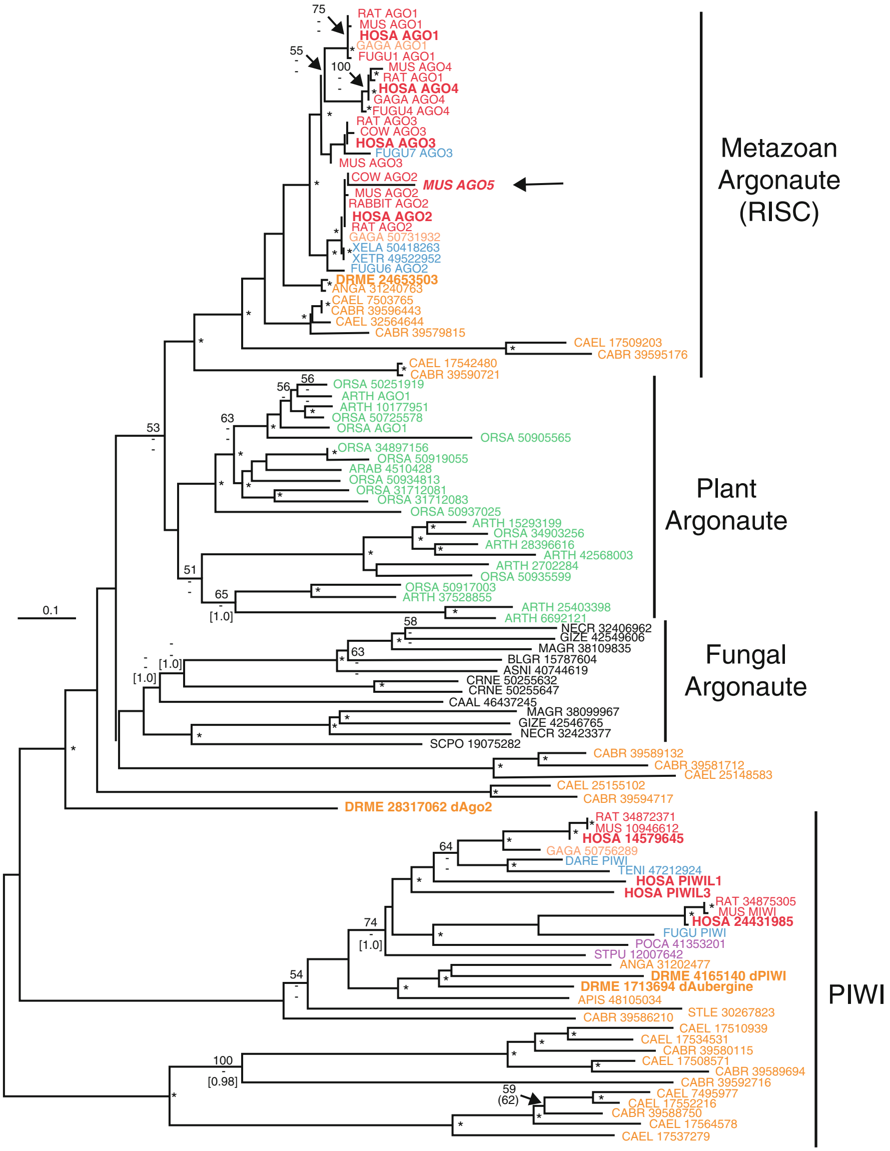


Fig. 3.9 Neighbor-joining phylogenetic tree of Argonaute/PIWI protein family. Major organism groups (with colours) are mammals (*red*), birds (*light red*), cold-blooded vertebrates (*deep blue*), urochordates (*light blue*), deutrostome invertebrates (*purple*), protostome invertebrates (*orange*), plants (*green*), fungi (*black*), and protists (*light purple*). The tree is unrooted. *Asterisks* ("*") indicate those nodes supported 60 % or greater occurrence and >0.90 probability in square parentheses. Nodes with one or two values less than 50 % have *dashes* ("–") while values less than 50 % are unmarked. *Scale bar* represents 0.1 expected amino acid residue substitutions per site. Human and Drosophila PIWI/Ago proteins are in larger font. The branch leading to a putative, but unlikely, fifth Argonaute gene homolog in mouse, mAgo5, is labeled with a large *arrow*. Other branches are labeled by a four letter species identifier (the first two letters from the genus and species names) and the GenBank accession number). From Murphy D, Dancis B, Brown JR. The evolution of core proteins involved in microRNA biogenesis. BMC Evol Biol. 2008; 8:92 (Open Access) [73]

The independent and lineage-specific evolution of the Argonaute family of genes may have resulted from the specific cell-line diversification in protein expression profiles, which could have evolved independently in each species because of specific developmental needs, i.e. the role of *Drosophila* Piwi in ovarian somatic cells vs. Aub and Ago3 involvement in TE cleavage in germ cells [50, 109–112]. The expansion of AGO and PIWI homologues and their functional specialization after gene duplication events, appears to have been well conserved during evolution, and can be observed even in distantly related organisms (Fig. 3.9) [73, 113, 114]. A representative example is the pea aphid, which expresses eight copies of *piwi* and two copies of *ago3* genes that have evolved separately in the germ and somatic cells of the organism, respectively, thus providing a greater plasticity for its sexual and asexual reproductive cycles [71]. Studies on the *Drosophila* piRISC assembly proteins have indicated that evolutionary selection dynamics would have favoured codon bias in the effort to combat TE abundance and invasions. That is, genes of the piRNA machinery with greater translational and possibly functional efficiency may have been selected through evolution in order to cope as best as possible with TE expansions [113].

Location of piRNA Biogenesis, and piRNA Origin and Evolution

Location of piRNA Biogenesis

Experimental evidence indicates that *Drosophila* piRNA biogenesis and the loading of mature piRNA onto PIWI proteins destined for nuclear function(s), occur in the germ plasma and in distinct cytoplasmic granules, proximal to the nucleus of germ cells. Similar granules make up the nuage structures around the nucleus of nurse oocyte cells, and the Yb-bodies of somatic ovarian cells [20, 115]. All of these granules are associated with active transcription, PGC specification and maintenance, and with germ cell differentiation and maturation. Granule formations require the hierarchical recruitment of components, such as developmental factors, TDRD proteins, and other components, a number of which form the macromolecular assembly for piRISC function(s) as, previously discussed. TDRD protein expression is gender and cell specific, and TDRD mutations affect negatively piRNA biogenesis, loading of mature piRNA to PIWI proteins and TE repression, which result in the abnormal development of germ cells and sterility [49, 69, 94, 98, 99].

Cytoplasmic granules similar to those observed in *Drosophila*, also exist in mammals. In the mouse male embryonic germ cells, Mili and Miwi2 are localized in two different granules (P-bodies). In the adult mouse testes, the P-body components form the inter-mitochondrial cement area seeing in spermatocytes, and later on, the chromatoid body seeing in round spermatids. The P-bodies appear to be the key sites for piRNA biogenesis, and to interact with each other functionally and physically during the critical developmental stage of *de novo* DNA methylation. They contain either Mili (called pi-bodies), or Miwi2 (called piP-bodies), and different TDRD proteins and other associate molecules required for the piRISC assembly structure, and for its localization and function(s) during gametogenesis. Disruption of the Mili pi-bodies affects negatively the integrity of piP-bodies but not *vice versa*, suggesting that Miwi2 is downstream of Mili regarding cytoplasmic granule communications. However, disruption of both types of P-bodies causes down regulation of piRNA production, accumulation of TEs and reduction in *de novo* DNA methylation, which result in germ cell loss and sterility [38, 39, 90, 91, 116]. Like in *Drosophila*, the integrity of P-bodies depends on the expression of TDRD and other proteins. Miwi is present in the whole of the cytoplasm in spermatocytes, and it concentrates in the chromatoid bodies in the round spermatids [81]. Mili and/or Miwi mutations, in combination with deficiencies in specific TDRDs, cause reductions in piRNA levels in mouse male germ cells, and increase retrotransposon transcription (LINE and LTR). These effects are associated with DNA hypomethylation, developmental germ cell defects and sterility [49, 85, 87, 88, 101–105].

Genomic Origin of piRNAs and Processing Selection

Deep sequencing technology has greatly aided in the characterization of the size and sequence of individual piRNAs [117–119]. It has also led to the development of algorithms to predict piRNA sequences and their genomic origin [119–122]. *Drosophila* piRNAs map mainly within intergenic pericentromeric and telomeric repetitive sequences including TEs [33, 34]. The majority of mammalian piRNAs map at intergenic, intronic and exonic sites, the functional significance of which is relatively unknown. Furthermore, although in *Drosophila* and Zebra fish most piRNAs have antisense homologies with transposon transcripts from all major classes of TEs, in mice only 17–20 % of piRNAs have any correspondence to known repetitive TEs and to retrotransposon coding regions. The remaining mammalian piRNAs (about 80%) correspond to unique non-coding sequences that represent the majority of “junk DNA” and may have important regulatory roles as new functions of “junk” DNA are being discovered [31–33, 43–47].

Sequence comparisons of piRNAs isolated from *in vitro* and *in vivo* studies, have shown that the hundreds of thousands of piRNAs are highly variable, and they are processed from long single stranded RNA precursors encoded in large tandem arrays of chromosomes, called clusters, ranging from 20 to 100 kb [47, 119, 120]. piRNA precursors may originate from one strand (+/–), called “uni-strand” or “monodirectional” clusters, or from both strands, called “dual strand” clusters. In some cases, in mice and in flies, the two halves of piRNA precursor may map to both genomic strands but in opposite orientations originating from “bidirectional clusters”, due to transcription from a centrally located promoter (0.1–1 kb long) [119].

In addition to the variability in the sequences of the piRNAs, there are differences in the frequency with which piRNAs map to specific cluster regions of chromosomes. piRNA densities of clusters can range from 40 to 4000 [120–124]. It has been debated as to whether the higher abundance of some piRNAs (i.e. those derived from intergenic non-coding sequences and from 3′ untranslated regions (3′ UTRs)), represent a programmed biogenesis of piRNAs with regulatory functions, or instead, local levels due to differences in translation efficiency, sequence motifs or stability. Robine et al. [124] argue that the 3′ UTR-derived piRNA production is, most likely, selected actively for *trans* regulatory functions since in *Drosophila* and mouse gonads, the piRNAs come from distinct piRNA-precursor categories. In terms of evolution, the co-transcription of piRNA clusters within protein coding gene sequences may have provided a greater range of regulatory piRNAs for a more efficient response to cell-specific requirements during the various stages of gametogenesis.

Master loci in *Drosophila*, express putative regulatory transcripts as substrates for the production of piRNAs that target protein mRNA coded in transposons. Examples include: the locus of the suppressor of the stellate gene (*su*(*ste*)), which prevents the over expression of stellate protein that crystallizes in spermatocytes and causes sterility [109, 125]; the *flamenco* locus involved in the repression of LTR retrotransposons of the *gypsy* family (*gypsy*, *ZAM* and *Idefix*), and the *3R-TAS* locus responsible for the silencing of the *P* element involved in female hybrid dysgenesis [16, 17, 21, 126]. Control of transposon transcription via piRISC in *Drosophila*, may have evolved after the transposition of a TE into a piRNA cluster. This event would have provided an evolutionary advantage to the fly by enabling it to discriminate mobile transposons from transcripts of endogenous genes, and eliminate them [50, 109, 111–114, 124].

The few studies on the mechanism of regulation of piRNA transcription from clusters in *Drosophila*, have indicated that uni-strand cluster transcription involves Polymerase II that requires a transcription start site and histone methyltransferase (H3K4me2) activity [127]. Transcription of dual-strand piRNA clusters in germ and somatic cells of the gonads, require instead the presence of double stranded DNA for the binding of specific epigenetic and transcription factors (i.e. HP1a, H3K9methyl-3, Cuff, Del). Binding is regulated by the heterochromatin Rhino, which acts as a licensing factor by distinguishing piRNA loci for cluster expression [127–129].

Taken together, it would appear that in *Drosophila*, the piRNA-PIWI pathway may regulate piRNA cluster expression or TGS by modifying chromatin status at specific genomic targets, via a piRISC-recruitment of epigenetic and transcription factors [125–130]. Heterochromatin binding of piRISC may involve direct piRNA binding to DNA, whereas euchromatin binding may involve binding to nascent RNA transcripts of 100–800 bp long, needed for the recruitment of histone methyltransferase and other proteins [130]. An H3K9 dependent vs. independent mechanism and the piRNA-Piwi complex bound to specific transcription factor(s) (i.e. the gametocyte-specific factor 1), have been shown to be involved in differentiating piRNA-precursor-transcription from piRNA-Piwi-repression of TE transcription [21, 130]. Specific RNA binding proteins may also be involved in the tagging and movement of cluster-derived RNA to the cytoplasm for the delivery of signals as to which PIWI protein may bind and process the precursor piRNA.

Comparisons on the similarities between rat and mouse genome, suggest that in mammals, most piRNA clusters may have originated via ectopic recombination and insertion of long sequences at regions flanked by chromosome-specific repetitive elements [117, 123]. Since in mice, there are no master loci as in *Drosophila*, potential regulatory sites for piRNA-cluster expression may most likely be interspersed in the genome and regulated by chromatin remodelling in a manner similar to that seen in *Drosophila*. Alternatively, expression of developmentally regulated mRNA transcripts may provide the substrates for the generation of piRNAs in response to specific need for TGS and/or PTGS. Present evidence indicates that the expression of specific RNA clusters during mouse spemiogenesis is regulated by zinc-type-transcription factors and the MYB-related protein A [131, 132].

piRNA Cluster Evolution

Comparisons of orthologous regions between rat, mouse and human genomes using the sequences of 140 known rodent piRNA clusters, have identified 37 of these to be of an ancestral origin and conserve in all three species. These clusters overlap protein-coding genes by spanning several exons and introns in most cases. The remaining 103 clusters are contained within intergenic regions and do not overlap protein-coding genes. Forty three of the one hundred and three clusters appear to have an ancestral origin, the remaining 60 appear to have evolved recently with most (44 clusters) having been acquired after rodent–primate divergence without a single cluster loss. Rapid expansion occurred before the rat-mouse split through recombination and long sequence insertions, primarily within genomic regions that were not preserved after cluster acquisition. In terms of large scale evolution, the rate of cluster expansion appears to exceed the highest known rates for any of the families of mammalians genes. On a small scale, however, the evolutionary rates of piRNAs are similar to those observed for other mammalian sequences. On given evidence, it has been suggested that piRNA cluster expansion is driven by positive selection, possibly as the result of the interplay between invading TEs and the transposition of existing TEs into clusters. Mammals in response to such evolutionary pressures may have selected to enhance the piRNA repertoire for silencing TEs [117, 125].

The mammalian piRNA clusters are by majority syntenic (i.e. their presence and chromosomal organization are conserved in the mouse, rat and human) [31–33, 43–47]. Positive selection and conservation of the piRNA cluster location, however, does not result in piRNA sequence conservation between closely related species and even between individual animals of the same species, as demonstrated from the sequencing of piRNAs isolated from the *Xenopous tropicalis* oocytes [133]. piRNAs are highly diverse due to the irregular manner of their biogenesis from different regions of cluster transcripts, and from clusters transcribed from one strand in preference to the other (strand asymmetry) [37, 51]. A quasi-random sub-saturation processing from common precursors has been suggested to explain the diversity in piRNAs and the low frequency of piRNAs with overlapping sequences [119]. Studies using computer simulations for TE transposition in *Drosophila*, and comparisons of sequences of a number of RNAi defence genes (including those of the piRISC) and of genes of the immune system, have suggested that intra-population and inter-species variations in transpositions, may contribute to the generation of individual piRNA sequences. Non-coding sequences are less conserved, and the generation of piRNAs from randomly cleaved long primary piRNA-precursors may further contribute to piRNA sequence diversity, as previously discussed. Based on the evidence available, it has been suggested that the piRNA defence system against transposons is most likely a dynamic one, and operates in a manner analogous to that of the immune defense system [113, 114, 134, 135]. A strong negative selection at the sequence level of piRNAs has only been observed in human African populations, in agreement with the findings that Africans have much higher rates of TE insertions than other populations [136, 137].

The Classes of piRNAs

The sequence signatures of piRNAs isolated from *Drosophila* and mice, have led to the proposal of two mechanistically different pathways for piRNAs biogenesis *in vivo* [31, 34, 43, 55, 65]: (a) The primary pathway that generates primary piRNA-precursors with U at the 5′ end, from long mRNA transcripts that are processed to mature piRNAs, as previously discussed. The biological significance of this in guiding/regulating piRISC or TGS is not known, at present. (b) The secondary pathway, also called the ping-pong amplification pathway, is assumed to use primary piRNAs to trigger the production of secondary piRNAs that drive TE-RNA degradation. Both pathways are active in the germline, whereas only the primary pathway is involved in piRNA biogenesis in gonad somatic cells [138].

piRNA isolation and sequencing have revealed some features that are shared between the diverse mature piRNAs species and are used to identify their origin and to understand the mechanism(s) of mature piRNA biogenesis [18, 21, 31–34, 43, 47]:

1. The mammalian piRNAs and *Drosophila* piRNAs/rasiRNAs as well as the pi-like-RNAs in other somatic cells/tissues [118], are larger (24–32 nt long) than miRNAs and siRNAs.
2. Mature piRNAs and primary piRNA-precursors carry a phosphate group at their 5′ end, indicative of being generated from a dsRNA precursor intermediate that is processed by an endonuclease.
3. Mature piRNAs are 2′-O-methylated at their 3′ end for greater stability [77, 79].
4. Mature piRNAs derived from piRNA-precursors via primary processing, show a very strong preference for U at the 5′ end and no nucleotide bias at position 10. Those derived via secondary processing show instead, a bias for Adenine (A) at position 10 and no 5′ end bias.
5. Antisense piRNAs derived from RNA transcripts containing TE repeats may share higher complementarity with sense piRNAs at the 5′ end [21, 31–34, 43–47].

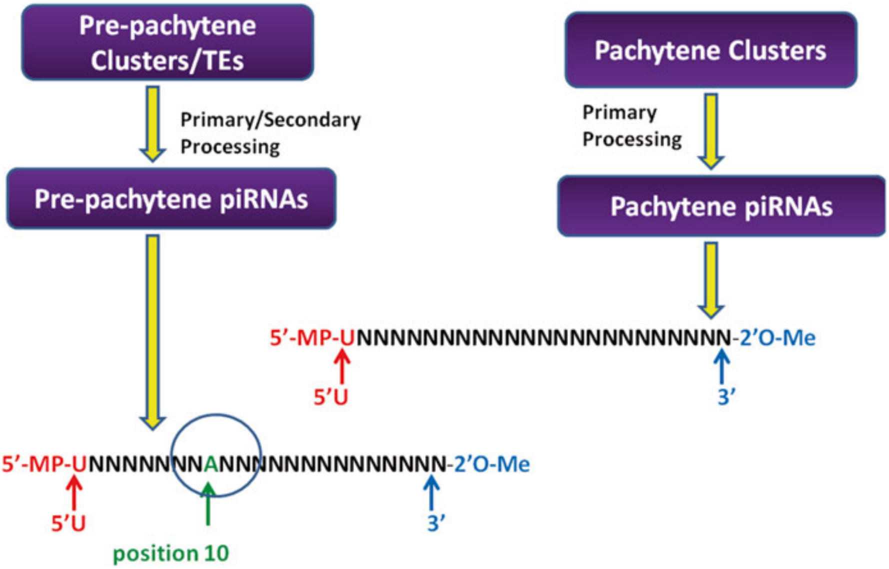

Pre-pachytene Clusters/TEs
Primary/Secondary Processing
Pre-pachytene piRNAs
Pachytene Clusters
Primary Processing
Pachytene piRNAs
5′-MP-UNNNNNNNNNNNNNNNNNNNNNNNNN-2′O-Me
5′U
3′
5′-MP-UNNNNNNNNANNNNNNNNNNNNNNN-2′O-Me
5′U
position 10
3′

Fig. 3.10 The major classes of mammalian piRNAs. *A* Adenine, *U* Uracil, *N* any ribonucleotide (Adenine, Uracil, Cytosine or Guanine)

Because there are no other known secondary structures motifs to distinguish piRNAs (i.e. hairpin-like structures as in miRNAs and siRNAs), their classification in mammals, is based on the above features, origin of piRNA cluster/strand and the stage of gametogenesis during which piRNAs are most abundant. The two major and distinct populations identified are the pre-pachytene and pachytene classes of piRNAs (Fig. 3.10).

The Pre-pachytene Class of piRNAs

This class of mammalian piRNAs, originate in a set of clusters (referred to as pre-pachytene piRNA clusters), and match those of *Drosophila* piRNAs originating in transposon master loci. These are "dual-strand clusters", rich in repetitive elements. The mouse pre-pachytene clusters are distinct, more numerous, and smaller in genomic coverage than the clusters of the pachytene class. Pre-pachytene piRNAs expressed in male growth-arrested gonocytes during embryonic development, are detected up to the mature spermatid-stage in adulthood. The majority show a sense bias: they contain A at position 10 (Fig. 3.10), and their first ten nucleotides are often complementary to piRNAs derived from antisense precursors. Pre-pachytene piRNAs are considered, therefore, to be processed from RNA transcripts via the ping-pong amplification pathway (secondary processing) [34, 50, 55, 61, 68, 69, 86].

Pre-pachytene piRNAs associate primarily with Miwi2 (80 %) and Mili, in mouse gonocytes, and with Mili and Miwi after birth and during adulthood. Their presence is related primarily to the *de novo* methylation of DNA and the silencing of TE via piRISC [55, 58–65]. Reduction in pre-pachytene piRNAs due to loss of Mili or Miwi activity, results in retrotransposon expression and male sterility, as previously discussed [63, 64, 83]. The piRNA-like small RNA (pilRNA) identified in hematopoietic cells (B-cells, T-cells, NK-cells), and in somatic cells from various tissues of other animals including flies, mice and rhesus macaque monkeys, resemble more the pre-pachytene class of piRNAs, and are considered to be involved in TE and/or endogenous retrovirus suppression [118].

The Pachytene Class of piRNAs

The mammalian pachytene class of piRNAs [34, 44, 47, 64, 65, 139, 140], are found in abundance (>90 %), in testicular germ cells during the pachytene stage, when spermatocytes enter meiosis to become spermatids. Pachytene piRNAs persist until the haploid round spermatid stage after which they disappear gradually during sperm differentiation. The majority (70 %), map to highly conserved monodirectional ("uni-strand") piRNA clusters, which are transcribed into long precursors ranging in length, from several to >100 kb. These clusters, are distributed over most chromosomes, mainly at intergenic regions but also at protein coding-genic regions (exons and introns) or 3′ UTRs, and they show no repeats or evidence for a transposon origin [44, 64, 124]. Pachytene piRNAs show primarily a 5′ end U bias, no bias for A at position 10, and no significant complementarily to each other (Fig. 3.10). They are considered therefore, to be the products of primary processing. As mentioned above, the functional significance of these piRNAs is unknown. They associate primarily with Miwi but also with Mili [64, 65, 83, 139]. It has been suggested that pachytene piRNAs may represent the degradation of end products of mRNA, which are destined for clearance during the final differentiation of spermatids into sperm [65].

Drosophila Classes of piRNAs

piRNAs from *Drosophila* germ line and nurse cells, have been mapped to discrete clusters transcribed from loci that contain defective transposons, or from master loci that contain active transposons [21, 25]. Most piRNA clusters are "dual-strand clusters" with antisense bias. piRNAs generated from sense precursor transcripts show no 5′ end U, have A at position 10 (typically Ago3-bound piRNAs), and are complementary to the first ten nucleotides of piRNAs generated from antisense precursor transcripts (Aub/Piwi-bound piRNAs). The majority of piRNAs in *Drosophila* germ line cells, therefore, are the products of TE cleavage, via secondary processing. Mature piRNAs of germ cells originating in clusters that do not contain TE sequences or in 3′ UTRs, are thought to be the products of primary processing by Piwi instead.

In ovarian somatic cells of *Drosophila*, which express only Piwi, piRNAs are mainly the products of primary processing that takes place in cytoplasmic granules, called the Yb bodies [112, 115]. These piRNAs show a bias for U at position 1. They originate mainly in the antisense strand of "uni-strand" clusters, and are destined for maternal transmission acting as *trans*-silencers of TE transposition during the early stages of embryogenesis [21, 25, 50, 57, 68, 138, 141, 142]. A few piRNAs derived from select master loci may have specific functions that are required for somatic cell maintenance [124]. Overall, piRNAs in *Drosophila,* can be categorized into (a) the germ line piRNAs (resembling the pre-pachytene class of mammalian piRNAs), which are the products of secondary processing, and (b) the ovarian somatic piRNAs (resembling the pachytene class of mammalian piRNAs), which are the products of primary processing [21, 50, 112].

Genomic studies in sea anemones and sponges that diverged before the emergence of bilateral animals (e.g. human, flies and worm), have provided evidence that both types of piRNA may have existed since the origin of metazoan [105, 109].

PIWI Role in piRNA Biogenesis

Primary piRNA Biogenesis

The existence of an operational primary piRNA pathway was originally proposed from studies on *Drosophila* ovarian somatic cell, which express only Piwi [57, 138, 141–143]. Processing of primary piRNA-precursors from clusters (>200 kb long) requires the Piwi association with ribosomal components in the cytoplasm and/or cytoplasmic granules, to produce piRNA from both genomic strands [124, 138, 141]. piRNA-precursors are loaded onto piRNA-bound Piwi (ovarian somatic cells), and/or piRNA-bound Aub (nurse and germ cells), with the help of cytoplasmic granule proteins (the nuclease Zuc and the putative helicases Armi and Yb), to be tailor-processed from the 3′ end into mature piRNAs of a size that is determined by the footprint of bound piRNA [57, 84, 109, 112, 138]. Mature piRNAs are finally loaded onto Piwi, which enters the nucleus for downstream target recognition and silencing. In germ line cells, Yb is not expressed and therefore, only Armi and Zuc are considered to be involved in the biogenesis of primary piRNAs destined to initiate the secondary pathway [138, 141]. The *armi* and *zuc* gene homologues are conserved in mice, and in the absence of the *armi* germ line orthologue, spermatogenesis is blocked and both Mili and Miwi2 lack bound piRNA [144].

The direct role of Piwi/Miwi2 in primary piRNA biogenesis is not clear. Mature piRNAs can still be produced even when *Drosophila* Piwi or mouse Miwi2 are mutated. This means that primary piRNA production in the germ line is carried out either by sequence independent endonucleases, or by an AGO member as mentioned above [60, 83]. In line with these observations, the localization of Piwi/Miwi2 in the nucleus and the distinct sizes of bound piRNAs in *Drosophila* and mice, suggested that the main roles of Piwi/Miwi2 may be in TGS guided by the loaded piRNAs, rather than in piRNA biogenesis per se [115]. Recent studies in *Drosophila* with Piwi deficient ovarian germ or somatic cells [21, 60], or with cultured ovarian somatic cells [21], or from inserting ectopic piRNA targets in the fly genome [130], have shown that the Piwi is guided by piRNA to euchromatin DNA targets, which are silenced via the Piwi recruitment of epigenetic factors that convert DNA to chromatin. A role in piRNA biogenesis to silence transcription of a locus may still be required for complete TE silencing, and may involve the chaperoning and export to the cytoplasmic sites of Piwi bound to nascent transcripts by other proteins [145].

Data sequences of mammalian pachytene piRNAs isolated from *in vivo* and *in vitro* studies, indicate that the majority of piRNAs (70–80 %) that bind Mili and Miwi, have signatures of primary piRNA biogenesis in evidence that this pathway is operating in mammals [18, 32, 44, 65, 96]. Since both of these proteins associate also with piRNAs that bear the signatures of secondary processing, it is possible that *in vivo*, the PIWI-symmetric dimethylation status defined by the bound piRNA may dictate (in association with the other components of the piRISC assembly), whether Miwi and Mili are involved in primary or secondary piRNA biogenesis. Primary piRNAs generated from mRNA transcripts from various species, appear to be evolutionarily more conserved and may be involved in gene regulation at more than one level of transcription and translation [64, 124, 140].

It has been argued whether the long piRNA-precursor molecules may have a biological function in directing DNA methylation and/or TGS during mitosis/meiosis. Recent analysis of the correlations between DNA methylation and the density of piRNAs at piRNA coding regions, in human and mouse chromosomes, has indicated a preferential methylation close to such regions up to 16 Mb long, suggesting that long precursor piRNA-PIWI complexes may epigenetically control the expression of non-coding regions, in a manner similar to X-chromosome silencing [146].

The Ping-Pong Circular Model

The putative ping-pong mechanism is assumed to enable piRNA amplification and transposons silencing to occur simultaneously during active transposon expression, which increase the efficiency of TE degradation via the production of antisense piRNAs. The mechanism is considered to be driven by the maternally inherited mature piRNAs during early embryogenesis in *Drosophila* germ cells, and possibly in mammalian oocytes. It involves the catalysis of 5′ cleavage of TE-RNA strands of opposite orientation. This is via reciprocal cycles of slicer activity of Aub/Piwi (bound to antisense piRNAs) and Ago3 (bound to sense piRNAs), which amplify the pool of complementary sense and antisense secondary piRNAs that feed into the cycle. The piRNAs that associate with Aub or Ago3, often overlap at their 5′ end by

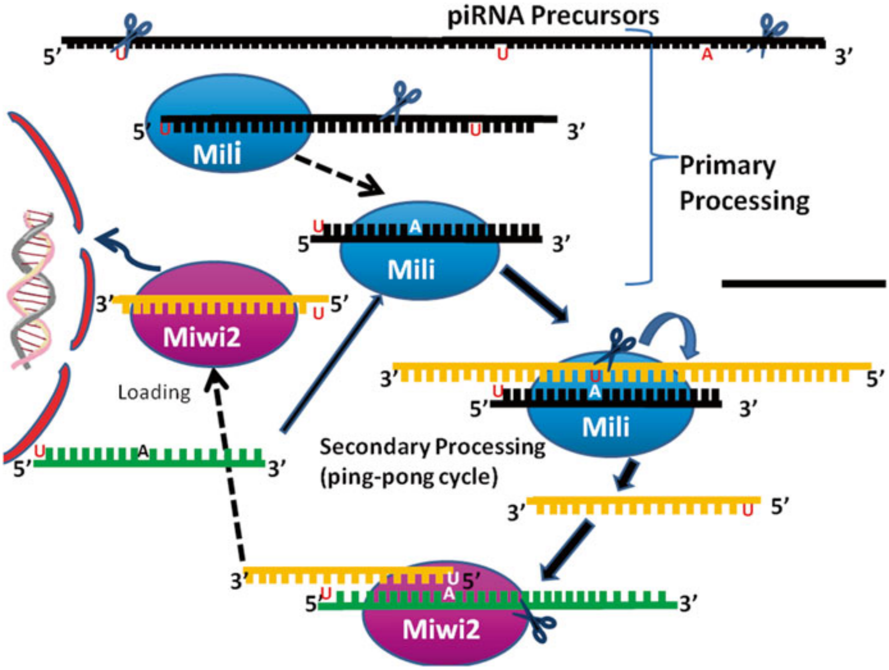


Fig. 3.11 The ping-pong model for piRNA biogenesis in mammals. Sense piRNAs derived from primary processing are used as guides by Mili to cleave antisense TE transcripts. The cleaved antisense transcripts serve in turn as guides for Miwi2 to cleave sense RNA thus fuelling an amplification cycle in which the 5′ ends of piRNAs are defined by PIWI cleavage. The 3′ ends are assumed to be shortened by an endonuclease and/or exonuclease and subsequently 2′-O-Methylated. Regulatory piRNAs may also be loaded onto to Miwi2 (*long dotted arrow*) for nuclear transport and piRISC function in TGS

ten nucleotides, which enables the transfer of the 5′-sliced precursor between the ping-pong partners that assemble on the surface of other proteins (i.e. vasa, TDRD proteins) [34, 64, 73, 138].

Based on the piRNA signatures generated, mice, unlike *Drosophila*, are assumed to use mature piRNAs derived from sense piRNA-precursors, in order to prime antisense piRNA production for TE-RNA degradation (Fig. 3.11). In foetal gonocytes, Mili associates primarily with sense piRNAs that have A at position 10, and Miwi2 associates with antisense piRNA with 5′ U, indicative of ping-pong signatures [19, 32, 34, 46, 51, 55]. However, studies with catalytically inactive Mili- and with Miwi2-DDH mutants, have shown that secondary piRNA biogenesis (pre-pachytene piRNA), and normal germ cell development depend on Mili and not on Miwi2 slicer-activity. Furthermore, in terms of piRNA levels, Mili appears to compensate for Miwi2 when the latter is not expressed [140]. piRNA analyses from *in vitro* (Mili) and from *in vivo* studies using Miwi- and Mili-DDH mouse mutants [64, 83, 103, 124], have shown that Miwi and Mili can catalyze secondary piRNA production (pre-pachytene piRNA), via an intra-amplification mechanism using complementary piRNAs as guides. As previously discussed, since the expression of Miwi2 but not its slicer-activity is required for *de novo* DNA methylation and normal gametogenesis, the role of Miwi2 may be limited to piRISC transcriptional and post-transcriptional functions [45, 83]. Alternatively, Miwi2 (like Mili and Miwi) may also promote primary piRNA biogenesis via the stabilization of mRNAs so that the mRNA can be cleaved by other endonucleases.

piRNA signatures for the function of the two pathways have been obtained from many animal species, including planarias, flies, zebra fish, frog, silkworm and mice, suggesting that they are conserved through evolution for the defence against invading TEs [33, 46, 51, 61, 66, 133, 147, 148]. Genetic studies on the evolutionary origin of the *Drosophila* somatic piRNA clusters and PIWI proteins, support the notion that the primary piRNA-Piwi pathway in these cells, evolved as a counter defence to the colonization of a follicular niche/soma by specific TEs, which tried to avoid germ line piRNA surveillance. Since mutations in genes that act only in the secondary pathway, result in the collapse of the entire piRNA pool, the role of the ping-pong mechanism could also be to ensure that all primary piRNA samples are used [23, 57, 110, 117, 134, 138, 141].

The Biological Significance of the piRNA-PIWI Pathway

Role in Human Diseases

A large number of investigations have provided evidence that the piRNA-PIWI pathway in metazoan has been associated with transcriptional and post transcriptional repression of TEs, and possibly of other genetic elements that are involved in the regulation of gametogenesis (from germ cell specification to GSC maintenance and proliferation, meiosis and to spermiogenesis). The biological significance of this pathway in human sperm development, and any relationship to idiopathic male sterility, however, have hardly been investigated. There has been one study in a Chinese population, on genetic variations in PIWI *vs.* spermiogenic failure. The study identified an SNP in the 3′ untranslated region of *HIWI2* associated with increased risk of oligozoospermia, and a non-synonymous SNP in *HIWI3* that was associated with reduced risk of oligozoospermia [149]. Some other studies have shown that the loss of the *HIWI* locus correlates with testicular atrophy and with the lack of development of proper secondary sexual characteristics [150, 151].

The effect of the *piwi* gene on germline development, has been related to the gonad somatic cell functions [40]. In lower organisms, it appears to play a role in genome rearrangement (Ciliates), synaptic plasticity and in associative memory development (Aplysia), as well as in whole-body regeneration (Colonial Ascidians) [132]. Maternal germline piRNA in *Drosophila,* has been implicated in the programming of somatic cells, and both maternal and zygotic Piwi proteins appear to be required for the establishment of heterochromatin and the suppression of phenotypic variations through epigenetic mechanisms. In addition, Piwi has been shown to have a role as an epigenetic activator [132]. A role of Piwi in somatic tissue differentiation, has been recently supported from studies on the time expression of piRNA during rat liver differentiation [152].

piRNA-PIWI It has been considered, therefore, that the "stem-like" proliferative phenotype of cancer cells could be related to the expression of components (i.e. PIWI proteins) involved in germ cell maintenance and self-renewal. Indeed investigations on PIWI re-expression in a cancer context, in humans, have shown that the Hiwi and Hili proteins and specific piRNAs (e.g. piRNA 651), are differentially expressed in precancerous and cancer cells depending on the cell type of the tumor (i.e. testicular (seminomas) and ovarian cancer, prostate, breast, gastrointestinal and endometrial cancers) [13, 22, 28, 104, 153–161].

Studies using ovarian tumours have shown that the high expression of *HIWI2* is associated with the silencing of genes that regulate apoptosis (i.e. Stat3/Bcl-X(L); p14ARF/p53), and with the expression of stem cell markers (e.g. *c-KIT*) thus correlating PIWI expression with the growth of cancer cells [13, 159–161]. The oncogen-like function of Hiwi2 was attributed to a Hiwi2-induced global DNA-hypermethylation of repetitive elements, which caused the silencing of tumour suppressor genes like the *CDKIs* [160]. Furthermore, Hiwi-mediated tumorigenesis was reversed by a methyltransferase inhibitor suggesting a connection between Hiwi2 and DNA methyltransferase [159].

Although investigations at the molecular level on the role of the piRNA-PIWI pathway in cancer are limited, the available studies suggest that if components of the piRNA-PIWI path are expressed in an aberrant way, they may contribute instead to the progression of cancer. In support, recent evidence has shown that the ectopic expression of piRNA pathway-associated genes contribute to the development of brain malignancies in *Drosophila* [168]. Furthermore, recent evidence has shown that promoter CpG island hypermethylation and inhibition of Piwil1, Piwil2, Piwil 4 and TDRD1 protein expression are related to reduced piRNA biogenesis and LINE-1 repetitive sequence hypomethylation, in primary seminoma and nonseminoma testicular tumors. Based on these, it is possible that the epigenetic inactivation of the PIWI-family of proteins could indeed contribute to male infertility and at least to the infertility-associated testicular cancer [162]. There are still many questions regarding the role of the PIWI proteins in cancer epigenetics. As matters stand at present, [153–161], the piRNA levels and the expression of specific piRNAs have the potential to serve as cell markers for cancer diagnosis, and/or to be used as treatment for altering the DNA methylation patterns to silence cancer related genes.

Conclusions

TEs have several opportunities to transpose themselves during gametogenesis, especially during periods when gene silencing and gene activation are taking place simultaneously (i.e. mitosis and meiosis of germ cells). The degradation of TEs transcripts via the direct catalytic activity of piRISC, and the repression of TE expression via piRISC-mediated epigenetic modifications, are two important functions that sexually reproducing organisms have evolved for defending TE invasions during gametogenesis. These functions have been well conserved and have evolved in mammals for optimal responses. Optimization may have driven the evolution of the spatiotemporal expression of PIWI proteins during male gametogenesis and the co-operation of multiple protein partners to facilitate piRISC activities in an orchestrated manner. This co-operation involves the receipt of messages (transcription of coding and non-coding mRNAs) during male gamete development and maturation, and the sorting and processing of these messages at specific cytoplasmic granules near the nucleus by the PIWI proteins and their associate partners. The generation of a large and highly diverse number of piRNAs that represent TE-RNA degradation products or serve as guides to degrade TEs, ensure feedback responses for piRISC and TGS so that the integrity of the genome is maintained [99]. Although the biological significance of piRNAs generated from non-coding intergenic DNA regions without repeats remains obscure at present, their number and diversity suggest that they may have functions beyond those of defending transposon mobilization. Based on studies in *Drosophila* and mice, such additional functions during gametogenesis include:

1. the degradation of mRNA (maternal) via recruitment of mRNA modifying enzymes [163];
2. the stabilization of coding mRNA via PIWI binding [31, 59, 63, 65] and,
3. the piRNA-PIWI-recruitment of transposable elements required for the assembly of the telomere protein complex to prevent telomere fusion [70, 164].

The piRNA-PIWI pathway may also have roles outside of the germ line as suggested by the evidence of the co-expression of PIWI-members and piRNAs in somatic stem cells of various tissues (including brain) of *Drosophila*, mouse and rhesus macaque [118] and in human brain [165]. Such roles are supported by the observe relationship between a specific piRNA and the repression of expression of the human melatonin receptor 1A gene [166], and the evidence for a piRNA regulatory role in the transcriptional regulation of plasticity-related genes involved in memory functions of the central nervous system of snails [167].

Compared with other small RNA systems, the piRNA-PIWI pathway is still not well understood, however. Many questions need to be answered with regard to the regulation/mechanisms of piRNA cluster transcription and primary piRNA biogenesis, the biological significance of the diverse piRNAs and the relationships of PIWI proteins with the various partners in cytoplasmic granules. The mechanisms and components that piRISC uses for the transcriptional/translational regulation of TEs and of other genes required for gametogenesis are been revealed slowly, and there are still a lot of questions.

Studies on the piRNA sequence and piRNA cluster evolution are consistent with the fast rates of evolution of the PIWI proteins in agreement with their interdependent roles in genome defence against transposons. The abundance and redundancy of mature piRNAs generated (with high sequence diversity), may also be part of the evolutionary strategy of organisms for coping with the fast rates of the ever expanding transposons. Whether different evolutionary pressures apply to piRNAs used for the regulation of transcription/translation of other genetic elements remain to be seen.

References

1. Biémont C (2010) A brief history of the status of transposable elements: from junk DNA to major players. Evolution 186:1085–1093
2. Zamudio N, Bourc'his D (2010) Transposable elements in the mammalian germ line: a comfortable niche or a deadly trap? Heredity 105:92–104
3. Hua-Van A, Le Rouzic A, Boutin TS, Filee J, Capy P (2011) The struggle for life of the genome's selfish architects. Biol Direct 6:19
4. Murr R (2010) Interplay between different epigenetic modifications and mechanisms. Adv Genet 70:101–141
5. Ballestar E (2011) An introduction to epigenetics. Adv Exp Med Biol 711:1–11
6. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13:484–492
7. Cui I, Cui H (2010) Antisense RNAs and epigenetic regulation. Epigenomics 2:139–150
8. Malecová B, Morris KV (2010) Transcriptional gene silencing mediated by non-coding RNAs. Curr Opin Mol Ther 12:214–222
9. Zhou H, Hu H, Lai M (2010) Non-coding RNAs and their epigenetic regulatory mechanisms. Biol Cell 102:645–655
10. Zamore PD, Tuschl T, Sharp PA, Bartel DP (2000) RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell 101(1):25–33
11. Kooter JM, Matzke MA, Meyer P (1999) Listening to the silent genes: transgene silencing, gene regulation and pathogen control. Trends Plant Sci 4:340–347
12. Hartig JV, Tomari Y, Förstemann K (2007) piRNAs-the ancient hunters of genome invaders. Genes Dev 21:1707–1713
13. Su C, Ren ZJ, Wang F, Liu M, Li X, Tang H (2012) PIWIL4 regulates cervical cancer cell line growth and is involved in down-regulating the expression of p14ARF and p53. FEBS Lett 586:1356–1362
14. Wang G, Reinke V (2008) A *C. Elegans* Piwi, PRG$_{-1}$, regulates 21U-RNAs during spermatogenesis. Curr Biol 18:861–867
15. Kim VN, Han J, Siomi MC (2009) Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol 10:126–139
16. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10:94–108
17. Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136(4): 656–668
18. Girard A, Hannon GJ (2008) Conserved themes in small-RNA-mediated transposon control. Trends Cell Biol 18:136–148
19. Lau NC, Robine N, Martin R, Chung WJ, Niki Y, Berezikov E, Lai EC (2009) Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome Res 19:1776–1785
20. Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12:246–258
21. Rozhkov NV, Hammell M, Hannon GJ (2013) Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev 27:400–412
22. Juliano C, Wang J, Lin H (2011) Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. Annu Rev Genet 45:447–469
23. Pillai RS, Chuma S (2012) piRNAs and their involvement in male germline development in mice. Dev Growth Differ 54(1):78–92
24. Simonelig M (2011) Developmental functions of piRNAs and transposable elements: a Drosophila point-of-view. RNA Biol 8:754–759
25. Khurana JS, Theurkauf W (2010) piRNAs, transposon silencing, and Drosophila germline development. J Cell Biol 191:905–913

26. Kota SK, Feil R (2010) Epigenetic transitions in germ cell development and meiosis. Dev Cell 19:675–686
27. De Felici M (2011) Nuclear reprogramming in mouse primordial germ cells: epigenetic contribution. Stem Cells Int 2011:e425863
28. Lee TL, Pang AL, Rennert OM, Chan WY (2009) Genomic landscape of developing male germ cells. Birth Defects Res C Embryo Today 87:43–63
29. Smallwood SA, Kelsey G (2012) De novo DNA methylation: a germ cell perspective. Trends Genet 28:33–42
30. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11:204–220
31. Grivna ST, Beyret E, Wang Z, Lin H (2006) A novel class of small RNAs in mouse spermatogenic cells. Genes Dev 20:1709–1714
32. Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764
33. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128:1089–1103
34. Aravin AA, Hannon GJ (2008) Small RNA silencing pathways in germ and stem cells. Cold Spring Harb Symp Quant Biol 73:283–290
35. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313:320–324
36. Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC (2006) Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev 20:2214–2222
37. Klattenhoff C, Theurkauf W (2008) Biogenesis and germ line functions of piRNAs. Development 135:3–9
38. Wang J, Saxe JP, Tanaka T, Chuma S, Lin H (2009) Mili interacts with Tudor domain-containing protein 1 in regulating spermatogenesis. Curr Biol 19:640–644
39. Aravin AA, van der Heijden GW, Castaneda J, Vagin VV, Hannon GJ, Bortvin A (2009) Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. PLoS Genet 5:e1000764
40. Cox DN, Chao A, Baker J, Chang L, Qiao D, Lin H (1998) A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes Dev 12:3715–3727
41. Kuramochi-Miyagawa S, Kimura T, Yomogida K, Kuroiwa A, Tadokoro Y, Fujita Y, Sato M, Matsuda Y, Nakano T (2001) Two mouse piwi-related genes: miwi and mili. Mech Dev 108:121–133
42. Deng W, Lin H (2002) miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell 2:819–830
43. Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germ line-specific class of small RNAs binds mammalian Piwi proteins. Nature 442:199–202
44. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE (2006) Characterization of the piRNA complex from rat testes. Science 313:363–367
45. Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ (2007) MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell 12:503–514
46. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, Plasterk RH, Hannon GJ, Draper BW, Ketting RF (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell 129:69–82
47. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, Chien M, Russo JJ, Ju J, Sheridan R, Sander C, Zavolan M, Tuschl T (2006) A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442:203–207

48. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316:744–747
49. Ku H-Y, Lin H (2014) PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. Natl Sci Rev 1:205–218
50. Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ (2009) Specialized piRNA pathways act in germ line and somatic tissues of the drosophila ovary. Cell 137:522–535
51. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455:1193–1197
52. Ohnishi Y, Totoki Y, Toyoda A, Watanabe T, Yamamoto Y, Tokunaga K, Sakaki Y, Sasaki H, Hohjoh H (2010) Small RNA class transition from siRNA/piRNA to miRNA during pre-implantation mouse development. Nucleic Acids Res 38(15):5141–5151
53. Gonzalez G, Behringer RR (2009) Dicer is required for female reproductive tract development and fertility in the mouse. Mol Reprod Dev 76:678–688
54. Cox DN, Chao A, Lin H (2000) piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development 127:503–514
55. Thompson T, Lin H (2009) The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annu Rev Cell Dev Biol 25:355–376
56. Gagnon KT, Corey DR (2012) Argonaute and the nuclear RNAs: new pathways for RNA-mediated control of gene expression. Nucleic Acid Ther 22:3–16
57. Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, Seitz H, Horwich MD, Syrzycka M, Honda BM, Kittler EL, Zapp ML, Klattenhoff C, Schulz N, Theurkauf WE, Weng Z, Zamore PD (2009) Collapse of germ line piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell 137:509–521
58. Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, Hata K, Li E, Matsuda Y, Kimura T, Okabe M, Sakaki Y, Sasaki H, Nakano T (2008) DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev 22:908–917
59. Unhavaithaya Y, Hao Y, Beyret E, Yin H, Kuramochi-Miyagawa S, Nakano T, Lin H (2009) MILI, a PIWI-interacting RNA-binding protein, is required for germ line stem cell self-renewal and appears to positively regulate translation. J Biol Chem 284:6507–6519
60. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Toth KF (2013) Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev 27:390–399
61. Aravin AA, Bourc'his D (2008) Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev 22:970–975
62. Watanabe T, Tomizawa S, Mitsuya K, Totoki Y, Yamamoto Y, Kuramochi-Miyagawa S, Iida N, Hoki Y, Murphy PJ, Toyoda A, Gotoh K, Hiura H, Arima T, Fujiyama A, Sado T, Shibata T, Nakano T, Lin H, Ichiyanagi K, Soloway PD, Sasaki H (2011) Role for piRNAs and non-coding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. Science 332:848–852
63. Reuter M, Berninger P, Chuma S, Shah H, Hosokawa M, Funaya C, Antony C, Sachidanandam R, Pillai RS (2011) Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature 480:264–267
64. Beyret E, Lin H (2011) Pinpointing the expression of piRNAs and function of the PIWI protein subfamily during spermatogenesis in the mouse. Dev Biol 355:215–226
65. Vourekas A, Zheng Q, Alexiou P, Maragkakis M, Kirino Y, Gregory BD, Mourelatos Z (2012) Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nat Struct Mol Biol 19:773–781
66. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC (2007) A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila. Science 315:1587–1590

67. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 453:534–538
68. Suh N, Blelloch R (2011) Small RNAs in early mammalian development: from gametes to gastrulation. Development 138(9):1653–1661
69. Lau NC (2010) Small RNAs in the animal gonad: guarding genomes and guiding development. Int J Biochem Cell Biol 42:1334–1347
70. Yin H, Lin H (2007) An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. Nature 450:304–308
71. Lu HL, Tanguy S, Rispe C, Gauthier JP, Walsh T, Gordon K, Edwards O, Tagu D, Chang CC, Jaubert-Possamai S (2011) Expansion of genes encoding piRNA-associated argonaute proteins in the pea aphid: diversification of expression profiles in different plastic morphs. PLoS One 6:e28051
72. Sasaki T, Shiohama A, Minoshima S, Shimizu N (2003) Identification of eight members of the Argonaute family in the human genome small star, filled. Genomics 82:323–330
73. Murphy D, Dancis B, Brown JR (2008) The evolution of core proteins involved in microRNA biogenesis. BMC Evol Biol 8:92
74. Qiao D, Zeeman AM, Deng W, Looijenga LH, Lin H (2002) Molecular characterization of hiwi, a human member of the piwi gene family whose over expression is correlated to seminomas. Oncogene 21:3988–3999
75. Sugimoto K, Kage H, Aki N, Sano A, Kitagawa H, Nagase T, Yatomi Y, Ohishi N, Takai D (2007) The induction of H3K9 methylation by PIWIL4 at the p16Ink4a locus. Biochem Biophys Res Commun 359:497–502
76. Jinek M, Doudna JA (2009) A three-dimensional view of the molecular machinery of RNA interference. Nature 457:405–412
77. Kirino Y, Mourelatos Z (2007) The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. RNA 13:1397–1401
78. Parker JS, Parizotto EA, Wang M, Roe SM, Barford D (2009) Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. Mol Cell 33:204–214
79. Tian Y, Simanshu DK, Ma JB, Patel DJ (2011) Structural basis for piRNA 2′-O-methylated 3′-end recognition by Piwi PAZ (Piwi/Argonaute/Zwille) domains. Proc Natl Acad Sci U S A 108:903–910
80. Nowotny M, Yang W (2009) Structural and functional modules in RNA interference. Curr Opin Struct Biol 19:286–293
81. Nguyen-Chi M, Morello D (2011) RNA-binding proteins, RNA granules, and gametes: is unity strength? Reproduction 142:803–817. doi:10.1530/REP-11-0257
82. Tan GS, Garchow BG, Liu X, Metzler D, Kiriakidou M (2011) Clarifying mammalian RISC assembly in vitro. BMC Mol Biol 12:19
83. De Fazio S, Bartonicek N, Di Giacomo M, Abreu-Goodger C, Sankar A, Funaya C, Antony C, Moreira PN, Enright AJ, O'Carroll D (2011) The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature 480:259–263
84. Haase AD, Fenoglio S, Muerdter F, Guzzardo PM, Czech B, Pappin DJ, Chen C, Gordon A, Hannon GJ (2010) Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila. Genes Dev 24(22):2499–2504
85. Sato K, Mishida KM, Shibuya A, Siomi MC, Siomi H (2011) Maelstrom coordinates microtubule organization during *Drosophila* oogenesis through interaction with components of the MTOC. Genes Dev 25:2361–2373
86. Siomi MC, Miyoshi T, Siomi H (2010) piRNA-mediated silencing in Drosophila germlines. Semin Cell Dev Biol 21:754–759
87. Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Takamatsu K, Chuma S, Kojima-Kita K, Shiromoto Y, Asada N, Toyoda A, Fujiyama A, Totoki Y, Shibata T, Kimura T, Nakatsuji N, Noce T, Sasaki H, Nakano T (2010) MVH in piRNA processing and gene silencing of retrotransposons. Genes Dev 24:887–892

88. Watanabe T, Chuma S, Yamamoto Y, Kuramochi-Miyagawa S, Totoki Y, Toyoda A, Hoki Y, Fujiyama A, Shibata T, Sado T, Noce T, Nakano T, Nakatsuji N, Lin H, Sasaki H (2011) MITOPLD is a mitochondrial protein essential for nuage formation and piRNA biogenesis in the mouse germline. Dev Cell 20:364–375
89. Zamparini AL, Davis MY, Malone CD, Vieira E, Zavadil J, Sachidanandam R, Hannon GJ, Lehmann R (2011) Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in Drosophila. Development 138:4039–4050
90. Chen C, Nott TJ, Jin J, Pawson T (2011) Deciphering arginine methylation: Tudor tells the tale. Nat Rev Mol Cell Biol 12:629–642
91. Vagin VV, Hannon GJ, Aravin AA (2009) Arginine methylation as a molecular signature of the Piwi small RNA pathway. Cell Cycle 8:4003–4004
92. Kirino Y, Kim N, de Planell-Saguer M, Khandros E, Chiorean S, Klein PS, Rigoutsos I, Jongens TA, Mourelatos Z (2009) Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. Nat Cell Biol 11:652–658
93. Girard A, Sachidanandam R, Hannon GJ, Aravin AA (2009) Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. Genes Dev 23:1749–1762
94. Liu K, Chen C, Guo Y, Lam R, Bian C, Xu C, Zhao DY, Jin J, MacKanzie F, Pawson T, Min J (2010) Structural basis for recognition of arginine methylated Piwi proteins by the extended Tudor domain. Proc Natl Acad Sci U S A 107:18398–18403
95. Kirino Y, Vourekas A, Sayed N, de Lima Alves F, Thomson T, Lasko P, Rappsilber J, Jongens TA, Mourelatos Z (2010) Arginine methylation of Aubergine mediates Tudor binding and germ plasm localization. RNA 16:70–78
96. Vourekas A, Kirino Y, Mourelatos Z (2010) Elective affinities: a Tudor-Aubergine tale of germline partnership. Genes Dev 24:1963–1966
97. Patil VS, Kai T (2010) Repression of retroelements in Drosophila germline via piRNA pathway by the Tudor domain protein Tejas. Curr Biol 20:724–730
98. Nagao A, Sato K, Nishida KM, Siomi H, Siomi MC (2011) Gender-specific hierarchy in Nuage localization of PIWI-interacting RNA factors in Drosophila. Front Genet 2:55
99. Ishizu H, Nagao A, Siomi H (2011) Gatekeepers for Piwi-piRNA complexes to enter the nucleus. Curr Opin Genet Dev 21:484–490
100. Chen C, Jin J, James DA, Adams-Cioaba MA, Park JG, Guo Y, Tenaglia E, Xu C, Gish G, Min J, Pawson T (2009) Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. Proc Natl Acad Sci 106:20336–20341
101. Siomi MC, Mannen T, Siomi H (2010) How does the royal family of Tudor rule the PIWI-interacting RNA pathway? Genes Dev 24:636–646
102. Shoji M, Tanaka T, Hosokawa M, Reuter M, Stark A, Kato Y, Kondoh G, Okawa K, Chujo T, Suzuki T, Hata K, Martin SL, Noce T, Kuramochi-Miyagawa S, Nakano T, Sasaki H, Pillai RS, Nakatsuji N, Chuma S (2009) The TDRD9-MIWI2 complex is essential for piRNA-mediated retrotransposon silencing in the mouse male germ line. Dev Cell 17:775–787
103. Reuter M, Chuma S, Tanaka T, Franz T, Stark A, Pillai RS (2009) Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile. Nat Struct Mol Biol 16:639–646
104. Hou Y, Yuan J, Zhou X, Fu X, Cheng H, Zhou R (2012) DNA demethylation and USF regulate the meiosis-specific expression of the mouse Miwi. PLoS Genet 8:e1002716
105. Cerutti H, Casas-Mollano JA (2006) On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet 50:81–99
106. Makarova KS, Wolf YI, van der Oost J, Koonin EV (2009) Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol Direct 4:29
107. Funayama N, Nakatsukasa M, Mohri K, Masuda Y, Agata K (2010) Piwi expression in archeocytes and choanocytes in demosponges: insights into the stem cell system in demosponges. Evol Dev 12:275–287

108. Kerner P, Degnan SM, Marchand L, Degnan BM, Vervoort M (2011) Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge Amphimedon queenslandica. Mol Biol Evol 28:2289–2303
109. Malone CD, Hannon GJ (2009) Molecular evolution of piRNA and transposon control pathways in Drosophila. Cold Spring Harb Symp Quant Biol 74:225–234
110. Carmell MA, Xuan Z, Zhang MQ, Hannon GJ (2002) The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev 16:2733–2742
111. Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. Cell 136:642–655
112. Olivieri D, Senti KA, Subramanian S, Sachidanandam R, Brennecke J (2012) The cochaperone shutdown defines a group of biogenesis factors essential for all piRNA populations in Drosophila. Mol Cell 47(6):954–969
113. Lu J, Clark A (2010) Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. Genome Res 20:212–227
114. Castillo DM, Mell JC, Box KS, Blumenstiel JP (2011) Molecular evolution under increasing transposable element burden in Drosophila: a speed limit on the evolutionary arms race. BMC Evol Biol 11:258
115. Qi H, Watanabe T, Ku HY, Liu N, Zhong M, Lin H (2011) The Yb body, a major site for Piwi-associated RNA biogenesis and a gateway for Piwi expression and transport to the nucleus in somatic cells. J Biol Chem 286:3789–3797
116. Vasileva A, Tiedau D, Firooznia A, Muller-Reichert T, Jessberger R (2009) Tdrd6 is required for spermiogenesis, chromatoid body architecture, and regulation of miRNA expression. Curr Biol 19:630–639
117. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10:135–151
118. Yan Z, Hu HY, Jiang X, Maierhofer V, Neb E, He L, Hu Y, Hu H, Li N, Chen W, Khaitovich P (2011) Widespread expression of piRNA-like molecules in somatic tissues. Nucleic Acids Res 39:6596–6607
119. Betel D, Sheridan R, Marks DS, Sander C (2007) Computational analysis of mouse piRNA sequence and biogenesis. PLoS Comput Biol 3:e222
120. Lakshmi S, Agrawal S (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res 36(Database issue):D173–D177
121. Zhang Y, Wang X, Kang L (2011) A k-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics 27:771–776
122. Rosenkranz D, Zischler H (2012) proTRAC – a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics 13:5
123. Assis R, Kondrashov AS (2009) Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. Proc Natl Acad Sci U S A 106:7079–7082
124. Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower MD, Lai EC (2009) A broadly conserved pathway generates 3′ UTR-directed primary piRNAs. Curr Biol 19:2066–2076
125. Kotelnikov RN, Klenov MS, Rozovsky YM, Olenina LV, Kibanov MV, Gvozdev VA (2009) Peculiarities of piRNA-mediated post-transcriptional silencing of stellate repeats in testes of Drosophila melanogaster. Nucleic Acids Res 37:3254–3263
126. Jensen PA, Stuart JR, Goodpaster MP, Goodman JW, Simmons MJ (2008) Cytotype regulation of *P* transposable elements in *Drosophila melanogaster*: repressor polypeptides or piRNAs? Genetics 179:1785–1793
127. Mohn F, Sienski G, Handler D, Brennecke J (2014) The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in *Drosophila*. Cell 157:1364–1379
128. Zhang Z, Wang J, Schultz N, Shang F, Parhad SS, Tu S, Vreven T, Zamore PD, Weng Z, Theurkauf WE (2014) The HP1 homolog rhino anchors a nuclear complex that suppresses piRNA precursor splicing. Cell 157:1353–1363

129. Pane A, Jiang P, Zhao DY, Singh M, Schupbach T (2011) The cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germ line. EMBO J 30:4601–4615
130. Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H (2013) A major epigenetic programming mechanism guided by piRNAs. Dev Cell 24:502–516
131. Ergin B (2009) Function of the mouse PIWI Proteins and biogenesis of their piRNAs in the male germline. Thesis, Duke University. http://hdl.handle.net/10161/1583
132. Ross RJ, Weiner MM, Lin H (2014) PIWI proteins and PIWI-interacting RNAs in the soma. Nature 505:353–359
133. Lau NC, Ohsumi T, Borowsky M, Kingston RE, Blower MD (2009) Systematic and single cell analysis of Xenopus Piwi-interacting RNAs and Xiwi. EMBO J 28:2945–2958, Erratum in: EMBO J. 2009; 28:3458
134. Kolaczkowski B, Hupalo DN, Kern AD (2011) Recurrent adaptation in RNA interference genes across the Drosophila phylogeny. Mol Biol Evol 28:1033–1042
135. Obbard DJ, Welch JJ, Kim KW, Jiggins FM (2009) Quantifying adaptive evolution in the Drosophila immune system. PLoS Genet 5:e1000698
136. Lukic S, Chen K (2011) Human piRNAs are under selection in Africans and repress transposable elements. Mol Biol Evol 28:3061–3067
137. Ewing AD, Kazazian HH Jr (2011) Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res 21:985–990
138. Xiol J, Spinelli P, Laussmann MA, Homolka D, Yang Z, Cora E, Coute Y, Conn S, Kadlec J, Sachidanandam R, Kaksonen M, Cisack S, Pehrussi A, Pillai RS (2014) RNA clamping by Vasa assembles a piRNA amplifier complex on transposons transcripts. Cell 157:1696–1711
139. Pillai RS, Chuma S (2012) piRNAs and their involvement in male germline development in mice. Dev Growth Differ 54(1):78–92
140. Gan H, Lin X, Zhang Z, Zhang W, Liao S, Wang L, Han C (2011) piRNA profiling during specific stages of mouse spermatogenesis. RNA 17:1191–1203
141. Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC (2010) Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. Genes Dev 24:2493–2498
142. Pek JW, Kai T (2011) Non-coding RNAs enter mitosis: functions, conservation and implications. Cell Div 6:6
143. Lin H, Yin H (2008) A novel epigenetic mechanism in Drosophila somatic cells mediated by Piwi and piRNAs. Cold Spring Harb Symp Quant Biol 73:273–281
144. Zheng K, Xiol J, Reuter M, Eckardt S, Leu NA, McLaughlin KJ, Stark A, Sachidanandam R, Pillai RS, Wang PJ (2010) Mouse MOV10L1 associates with Piwi proteins and is an essential component of the Piwi-interacting RNA (piRNA) pathway. Proc Natl Acad Sci U S A 107:11841–11846
145. Tianfang Ge D, Zamore PD (2013) Small RNA-directed silencing: the fly finds its inner fission yeast? Curr Biol 23:318–320
146. Sigurdsson MI, Smith AV, Bjornsson HT, Jonsson JJ (2012) The distribution of a germline methylation marker suggests a regional mechanism of LINE-1 silencing by the piRNA-PIWI system. BMC Genet 13:31
147. Montgomery TA, Rim Y-S, Zhang C, Dowen RH, Phillips CM, Fischer SE, Ruvkun G (2012) PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog henn-1. PLoS Genet 8:e1002616
148. Friedländer MR, Adamidi C, Han T, Lebedeva S, Isenbarger TA, Hirst M, Marra M, Nusbaum C, Lee WL, Jenkin JC, Sánchez Alvarado A, Kim JK, Rajewsky N (2009) High-resolution profiling and discovery of planarian small RNAs. Proc Natl Acad Sci U S A 106:11546–11551
149. Gu A, Ji G, Shi X, Long Y, Xia Y, Song L, Wang S, Wang X (2010) Genetic variants in Piwi-interacting RNA pathway genes confer susceptibility to spermatogenic failure in a Chinese population. Hum Reprod 25(12):2955–2961

150. Angulo MA, Castro-Magana M, Sherman J, Collipp PJ, Milson J, Trunca C, Derenoncourt AN (1984) Endocrine abnormalities in a patient with partial trisomy 4q. J Med Genet 21:303–307
151. Sathya P, Tomkins DJ, Freeman V, Paes B, Nowaczyk MJ (1999) De novo deletion 12q: Report of a patient with 12q24.31q24.33 deletion. Am J Med Genet 84:116–119
152. Rizzo F, Hashim A, Marchese G, Ravo M, Tarallo R, Nassa G, Giurato G, Rinaldi A, Cordella A, Persico M, Sulas P, Perra A, Ledda-Columano GM, Columbano A, Wisz A (2014) Time regulation of p-element-induced Wimpy testis-interacting RNA exoressuib during rat liver regeneration. Hepatology. doi:10.1002/hep.27267
153. Dyce PW, Toms D, Li J (2010) Stem cells and germ cells: microRNA and gene expression signatures. Histol Histopathol 25:505–513
154. Juliano CE, Swartz SZ, Wessel GM (2010) A conserved germline multipotency program. Development 137:4113–4126
155. Wang Y, Liu Y, Shen X, Zhang X, Chen X, Yang C, Gao H (2012) The PIWI protein acts as a predictive marker for human gastric cancer. Int J Clin Exp Pathol 5:315–325
156. Cheng J, Guo JM, Xiao BX, Miao Y, Jiang Z, Zhou H, Li QN (2011) piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. Clin Chim Acta 412:1621–1625
157. Alié A, Leclère L, Jager H, Dayraud C, Chang P, Le Guyader H, Quéinnec E, Manuel M (2011) Somatic stem cells express Piwi and Vasa genes in an adult ctenophore: ancient association of “germ line genes” with stemness. Dev Biol 350:183–197
158. Hashimoto H, Sudo T, Mikami Y, Otani M, Takano M, Tsuda H, Itamochi H, Katabuchi H, Ito M, Nishimura R (2008) Germ cell specific protein VASA is over-expressed in epithelial ovarian cancer and disrupts DNA damage-induced G2 checkpoint. Gynecol Oncol 111:312–319
159. Siddiqi S, Matushansky I (2012) Piwis and piwi-interacting RNAs in the epigenetics of cancer. J Cell Biochem 113:373–380
160. Siddiqi S, Terry M, Matushansky I (2012) Hiwi mediated tumorigenesis is associated with DNA hypermethylation. PLoS One 7:e33711
161. Chen L, Shen R, Ye Y, Pu XA, Liu X, Duan W, Wen J, Zimmerer J, Wang Y, Liu Y, Lasky LC, Heerema NA, Perrotti D, Ozato K, Kuramochi-Miyagawa S, Nakano T, Yates AJ, Carson WE 3rd, Lin H, Barsky SH, Gao JX (2007) Precancerous stem cells have the potential for both benign and malignant differentiation. PLoS One 2:e293
162. Fereira HJ, Heyn H, Garcia del Muro X, Vidal A, Larriba S, Munoz C, Villanueva A, Esteller M (2014) Epigenetic loss of the PIWI/piRNA machinery in human testicular tumorigenesis. Epigenetics 9:113–118
163. Rouget C, Papin C, Boureux A, Meunier AC, Franco B, Robine N, Lai EC, Pelisson A, Simonelig M (2010) Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. Nature 467:1128–1132
164. Shpiz S, Olovnikov I, Sergeeva A, Lavrov S, Abramov Y, Savitsky M, Kalmykova A (2011) Mechanism of the piRNA-mediated silencing of Drosophila telomeric retrotransposons. Nucleic Acids Res 39:8703–8711
165. Lee E, Banerjee S, Zhou H, Jammalamadaka A, Arcila M, Manjunath BS, Kosik KS (2011) Identification of piRNAs in the central nervous system. RNA 17:1090–1099
166. Esposito T, Magliocca S, Formicola D, Gianfrancesco F (2011) piR_015520 belongs to Piwi-associated RNAs regulates expression of the human melatonin receptor 1A gene. PLoS One 6(7):e22727
167. Rajasethupathy P, Antonov I, Sheridan R, Frey S, Sander C, Tuschl T, Kandel ER (2012) A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. Cell 149:693–707
168. Mendzabal JA, Llamazares S, Rossell D, Gonzalez C (2010) Ectopic expression of germline genes drives malignant brain tumor growth in *Drosophila*. Science 330:1824–1827

Chapter 4
Long Non-coding RNA

Monika Gullerova

Introduction

Rapid developments in sequencing technologies during past decade revealed that protein coding genes represent only 2 % of the human genome [1]. This came across as a surprise considering that genomes of the other organisms (from yeast to *Caenorhabditis elegans*) are quite dense. Is the rest of the human genome merely "junk" DNA? This question was partially answered by experimental approaches like high-throughput sequencing or whole genome high density tailing arrays. Now it is known that this "junk" DNA is transcribed throughout mammalian genomes and because it lacks protein coding capacity it is referred to as long non-coding RNA (lncRNA) [2–7].

Non-coding transcripts have similar structure to messenger or coding RNA (mRNA): they are transcribed by RNA polymerase II (Pol II), they possess a cap and a polyA tail and they can be even spliced. However their function remains enigmatic [8]. Non-coding RNAs can be also divided into two groups based on their length: short and long non-coding RNAs (ncRNAs) or based on their primary function: structural and regulatory ncRNA (Fig. 4.1). Unlike mRNA and structural ncRNAs, most of lncRNAs are localized in nucleus [7].

Prior to the sequencing era, some lncRNA were discovered using old-fashioned gene cloning methods. Initially they were thought to be coding RNAs, however, deeper analyses revealed a lack of open reading frames (ORF). Furthermore they were thought to be random in nature more so than stable elements in genome. This opinion changed when FANTOM consortium analyzed over 60,000 full length cDNAs and identified over 11,000 lncRNAs in mouse [9]. Interestingly, a large portion of identified lncRNAs

M. Gullerova, Ph.D. (✉)
Sir William Dunn School of Pathology, University of Oxford,
South Parks Road, OX1 3RE Oxford, UK
e-mail: monika.gullerova@path.ox.ac.uk

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_4

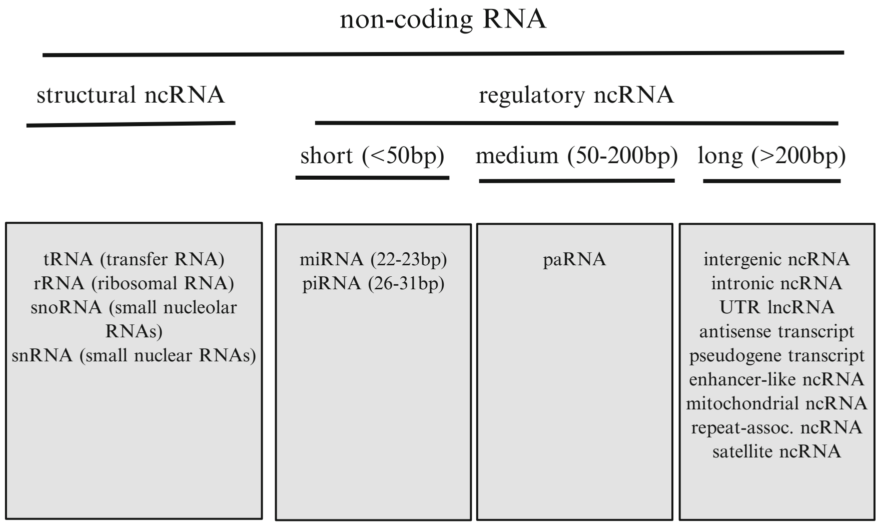


Fig. 4.1 Summary of various types of non-coding RNAs

is transcribed in antisense orientation to protein-coding genes, thus are referred to as natural antisense transcripts (NATs). Another study extended these findings by identifying NATs in human genome [10]. Interestingly many cancer associated genes, particularly tumor suppressor genes have long antisense ncRNA.

More recently, it was shown that intergenic regions also express thousands of long non-coding RNAs, named large intervening non-coding RNAs (lincRNAs). These transcripts were discovered by analyses of active chromatin marks (H3K4 acetylation and H3K36 trimethylation) genome-wide and eliminating those regions corresponding to protein coding genes and microRNAs. This approach was followed by extended analyses using RNA-seq experiments [4–6, 11–13]. Up to date, there are more than 8000 lincRNAs identified. More than half one them are confirmed lincRNAs; they can be localized in nucleus, cytoplasm or both and they are multi-exonic, capped and polyadenylated.

A major interest now lies in functional analysis of lncRNA. The fact that they are not evolutionary conserved, even between related species could indicate that most of them are non functional and may represent just transcriptional noise. To date, only 200 lncRNAs have been studied functionally with variable outcome, although many of them show at least functional evidence in vitro. Only a few lncRNAs have been studied in animal models, suggesting that they are not essential for viability. For example, mouse homologue of HOTAIR is poorly conserved in sequence and its deletion, along with the deletion of the HoxC cluster, has only a little effect in vivo, neither on the expression pattern or transcription efficiency, nor on the amount of K27me3 coverage of different Hoxd target genes [14]. On the other hand it is possible that the lack of one particular lncRNA can substituted by another one.

Overall, it is certain that mammalian genomes, including humans, produce thousands of lncRNAs. Due to the complexity and variability of lncRNAs it may take several years to understand and clarify their functional identity. It is very likely that some lncRNA are involved in a range of biological processes. In this chapter, I will discuss a number of important topics regarding lncRNAs: their origin, function, and role in disease.

Origins and Genome Localization of lncRNA

Low degree of lncRNA conservation among species suggests their different evolutionary design to protein coding genes. The limited phylogenetic range of lncRNA could be also explained by specific existence, but rapid declination within particular lineages [15]. A first possible scenario for lncRNA emergence is metamorphosis of protein-coding gene into non-coding RNA sequence. A protein-coding gene may under go mutations such as a frame shift that disrupts its open reading frame while maintaining the expression of the RNA transcript. For example, *Xist* gene encodes an lncRNA that is crucial for X chromosome inactivation. Recently, it's been shown that several exons and the promoter of *Xist* are derived from the protein-coding gene *Lnx3* that has acquired frame shift mutations during early mammalian evolution [16, 17]. It is possible that such a metamorphosis involved two steps: initial degeneration of the original sequence, followed by subsequent emergence of residual exons into newly formed Xist gene. Possibly these events occurred concurrently (Fig. 4.2a). A nother possibility includes chromosomal rearrangement, when two separate sequences are joined and together create an expressed non-coding sequence (Fig. 4.2b). One such example comes from the observation that a dog testis-derived non-coding RNA has arisen only recently following a lineage specific change. Also duplications in a non-coding RNA sequence could cause repeats, increasing the length of the transcript. Rare examples of duplicated lncRNA include Neat2 (mouse nuclear enriched abundant transcript) and mouse testis-derived lncRNA that are separate paralogous to non-exonic sequences elsewhere in the

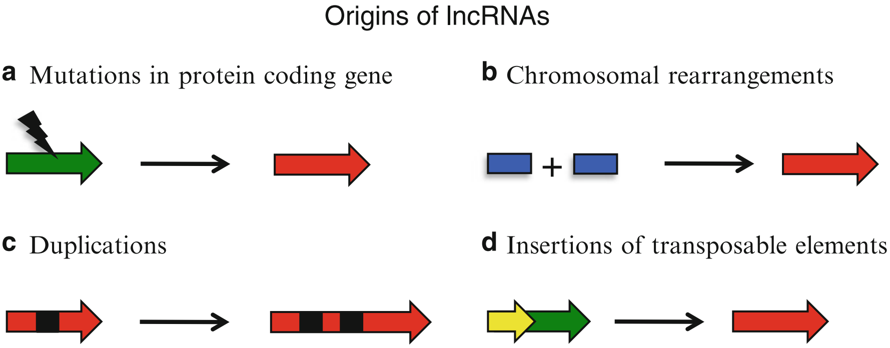

Origins of lncRNAs
a Mutations in protein coding gene
b Chromosomal rearrangements
+
c Duplications
d Insertions of transposable elements

Fig. 4.2 Origins of lncRNAs. *Block arrows* represent genes. *Black* regions in *block arrows* are introns. (**a**) Mutations in protein coding gene. (**b**) Chromosomal rearrangements. (**c**) Duplications. (**d**) Insertions of transposable elements

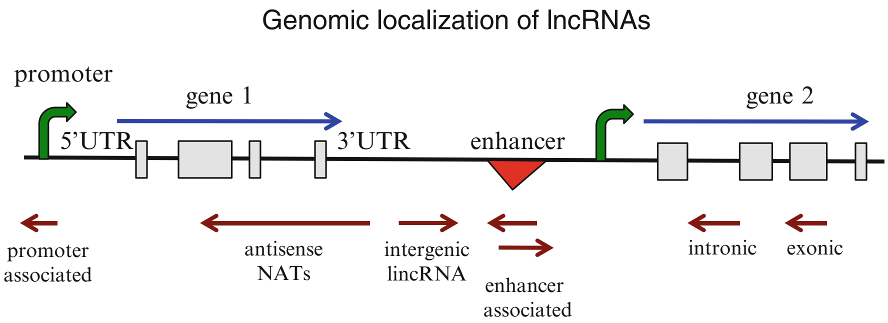


Fig. 4.3 Genomic localizations of lncRNAs. Schematic diagram illustrating organization of lncRNAs associated with protein-coding genes. *Arrows* pointing towards right represent sense transcription, *arrows* pointing towards left correspond to antisense transcription

genome [18]. Local, tandem duplication might also lead to generation of repeats increasing the size of lncRNA (Fig. 4.2c). Insertions of transposable elements may also emerge as lncRNA. For example, BC1 and BC200 (brain cytoplasmic RNA1 and 200-nucleotide) ncRNAs derive from two separate transposons, in the rodent and anthropoid lineages. Despite their lack of common origin, BC1 and BC200 are involved in similar roles in translational regulation. Furthermore, a transposable element containing a transcriptional start site can be inserted into the genome to create a functional, but noncoding RNA sequence [19] (Fig. 4.2d).

In respect of protein-coding genes, lncRNA can overlap with a gene as well as being associated with the gene's promoter region. They can be transcribed from intragenic sequence, exonic or intron, or from intergenic region (Fig. 4.3). In general, lncRNA are expressed at low levels and their expression varies with location, time, development and physiological stimuli.

Natural Antisense Transcripts (NATs)

Strand specific sequencing methods, like RNA-seq, reveal complex overlapping transcription with many lncRNA being transcribed from complementary DNA strands of protein coding genes. These are referred to as antisense transcripts (AS). AS can arise from novel transcription start sites as well as from bi-directional promoters, or through transcriptional read-through. In yeast, AS do not occur randomly, but have been linked to sexual differentiation or stress response genes [20–22] as well as genes with higher variability in expression. Furthermore, certain AS transcript pairs are conserved across several yeast species.

Recently genome-wide transcriptome studies reveal that natural AS transcripts (NATs) frequently occur across mammalian genomes [23, 24]. The sizes and features of NATs can be variable. They can be classified based on their transcriptional start site, splicing, capping and polyadenylation. With respect to transcriptional start sites, NATs can be divided into three types: (1) overlapping, (2) intronic and (3) intergenic [25, 26]. Sequence of overlapping NATs partly overlaps with sense mRNA sequence with preference for the 3′-untranslated region (3′ UTR). Due to sequence complementarity, these regions can mutually interact by complete or partial hybridization. Thus NATs may function to regulate the expression of the overlapping sense mRNA [27–29].

Long Intergenic ncRNAs (lincRNAs)

Mammalian genomes encode >1000 long intergenic noncoding (linc)RNAs that are clearly conserved across mammals and are potentially functional. These lincRNAs have been implicated in diverse biological processes, including cell-cycle regulation, immune surveillance, and embryonic stem cell pluripotency. However, the mechanism by which these lincRNAs function remains elusive. To date, there are approximately 3300 of human lincRNAs identified by analyzing chromatin-state maps in various human cell types. It is known that one of the well-characterized lincRNA HOTAIR binds the polycomb repressive complex (PRC2). Remarkably, approximately 20 % of lincRNAs expressed in various cell types are bound by PRC2. Other lincRNA interact with different chromatin-modifying complexes. Furthermore depletion of certain lincRNAs associated with PRC2 leads to changes in gene expression, causing the up-regulation of genes that are normally silenced by PRC2. Therefore it is suggested that some lincRNAs guide chromatin-modifying complexes to specific genomic loci to regulate gene expression [2, 5].

Promoter Associated ncRNAs (CUTs, PROMPTs)

Several genome-wide studies have revealed the unanticipated property of RNA polymerase II (Pol II) to initiate transcription in promoter regions in both directions. Such a bi-directional transcription results in so called cryptic unstable transcripts (CUTs) in budding yeast [30–32]. CUTs are Pol II-dependent transcripts produced from promoters in opposite direction to the coding gene, which are degraded by the nuclear exosome shortly after their synthesis. Another type of transcript derived from promoter regions are stable unannotated transcripts (SUTs), which are not processed by exosome [32]. Bidirectional transcription is not only limited to yeast species, but extends to higher eukaryotes too. Studies using exosome depletion in human fibroblasts have revealed lncRNAs, which correspond to upstream regions of protein coding genes. Such promoter-upstream transcripts are referred to as PROMPTs [33]. Furthermore, RNA-seq analysis from mouse embryonic stem cells has shown many promoter associated transcripts, which are transcribed in non-random, divergent orientation [34]. Another sequencing approach employed global run on sequencing (GRO-seq) to indentify nascent RNAs in human fibroblasts. This study revealed that almost 80 % of the active promoters display bidirectional transcriptional activity [35]. It is now known that bidirectional transcription is a widespread phenomenon, which is conserved across species. There are suggestions that this bidirectional transcription may be a type of gene expression regulation that promotes an open chromatin structure at promoters by recruiting positive or negative transcriptional regulators.

Enhancer Associated ncRNAs (eRNAs)

Enhancers act to regulate expression of protein-coding genes from a distance in an orientation independent manner. A recent study, which analysed genome wide location of the enhancer binding protein CBP, showed over 12,000 positive loci in mouse neurons. Further transcriptional analysis also confirmed that Pol II is present at 25 % of those loci. Enhancer associated transcripts were identified as lncRNAs (termed eRNAs) produced in both directions and their expression does correlate with mRNA synthesis from nearby gene [36]. There is a possibility that eRNAs may be directly involved in enhancer function. eRNAs may facilitate recruitment of enhancer-associated proteins or enhance chromatin looping to provide contact between the enhancer and the promoter of a particular regulated gene [37].

Repeat Associated ncRNAs

Retrotransposons are genetic elements that can amplify themselves within a genome. They are ubiquitous components of the DNA of many eukaryotic organisms, and are particularly abundant in plants, where they are often a principal component of nuclear DNA. In mammals, almost half the genome (45–48 %) comprises retrotransposons, which possess extensive transcriptional activity. FANTOM 4 project has revealed that in human and mouse genomes, retrotransposons are expressed in a tissue specific manner. They are located close to promoter regions of protein-coding genes, suggesting that they may play a role in controlling alternative promoters or in the post-transcriptional regulation of gene expression [38].

Pseudogenes are another type of repetitive elements that can be transcribed into lncRNA. These can regulate protein-coding genes through competition for regulatory miRNA binding [39, 40].

Biogenesis, Processing and Structure of lncRNAs

Biogenesis of lncRNA is very similar to mRNA. LncRNAs are produced from many regions across the genome by transcribing Pol II. Only some lncRNAs have been shown to be products of Pol III. The majority of lncRNAs are 5′ capped, polyadenylated and spliced. LncRNA can be divided into two categories based on their orientation: they can be encoded on positive or negative DNA strand (sense or antisense orientation).

Many lncRNA transcripts are not end products, but face to further processing into a final functional form. The presence of sense lncRNA, which contain exons from mRNA sequences and intronic lncRNA that are derived entirely from intronic sequence, has lead to the hypothesis that many lncRNA transcripts are unprocessed pre-mRNAs prior to splicing and that intronic lncRNAs are by-products of this processing step. However, this is not the case for all sense and intronic lncRNAs since the expression patterns of some of these transcripts are not the same as their associated protein-coding gene [41]. Another hypothesis suggests that some lncRNA sequences are precursors to short miRNAs. An example of this is the lncRNA *H19* which encodes the miRNA *miR-675* [4]. Based on this evidence, post-transcriptional processing may occur with many lncRNA transcripts, but until more of them are functionally defined this question remains open.

Secondary structure formation is an important consideration in lncRNAs because they are able to interact with proteins or genomic DNA via these structures. Recently, models used for prediction of secondary structure have redefined the question of lncRNA evolution by looking at sequence conservation or compensatory mutations that would maintain secondary structure motifs [42]. However, approaches that predict RNA secondary structures with high precision have yet to be developed. Therefore, the number of structured ncRNAs remains to be determined, but is expected in intergenic, intronic and UTR regions and lacking in exon sequence.

Mechanism of Action

Despite the fact that only a fraction of all identified lncRNAs has been examined experimentally, an emerging paradigm suggests that they are implicated in many biological contexts. To date, lncRNAs have been implicated in regulation of gene expression, guidance of chromatin-modifying complexes, X chromosome inactivation, genomic imprinting, nuclear compartmentalization, nuclear-cytoplasmic trafficking, RNA splicing and translational control [43–46].

Regulation of Chromatin Structure

LncRNAs have been implicated in epigenetic gene regulation. Recent studies propose two basic models for lncRNA action at the chromatin level:

1. epigenetic silencing in *cis*, where lncRNA transcripts coat gene clusters and silence their expression by making them inaccessible to transcription machinery. These lncRNAs can also recruit chromatin remodeling proteins to epigenetically mark the region for heritable gene silencing;
2. epigenetic silencing in *trans*: lncRNAs can interact with chromatin modifying proteins to epigenetically silence genes at another locus (Fig. 4.4).

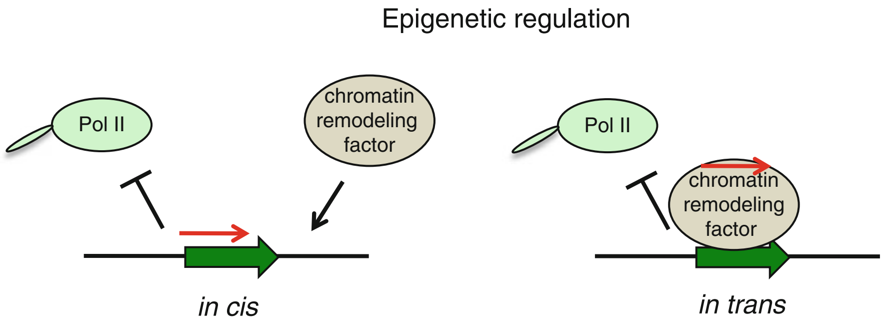


Fig. 4.4 Epigenetic regulation of gene expression by lncRNAs. *Block arrow* represents protein-coding gene, *red arrow* depicts lncRNA

Epigenetic Regulation in *cis*

One of the well studies lncRNAs to date is Xist, which is crucial for X chromosome inactivation in female somatic cells. It has been discovered in 1991 and despite of enormous effort, the exact mechanism of Xist-mediated X chromosome inactivation is still not fully understood. It is accepted that Xist is associated as an RNA compartment with the inactivated X chromosome [47]. The coating the chromatin that is silenced provides the first model for how lncRNA might function in stable epigenetic gene silencing in *cis*. Xist establishes a specialized, Pol II free, region, into which most of the X chromosome becomes localized during inactivation [48]. It should be noted that Xist coating of chromatin is stable even during metaphase, suggesting a form of epigenetic memory for the inactive X chromosome to remain silent over many cell divisions [49].

RepA, a small repeat region within Xist, is transcribed from both X chromosomes along with Tsix lncRNA (antisense partner of Xist) [50]. Tsix prevents RepA binding to either X chromosomes until post-cellular differentiation, when RepA in association with the chromatin-modifying complex PRC2 (polycomb repressive complex 2) binds to one of the two X chromosomes at the so called inactivation center [50]. Full-length Xist, produced from the X chromosome, destined to be inactivated, also binds to PRC2 and leads to the spreading of X inactivation from the center to the entire X chromosome in *cis*. Active chromatin status of the other X chromosome is protected by Tsix, which blocks transcription of Xist. It is not well understood, what prevents Xist from escaping the inactive X chromosome and acting on the active X chromosome in *trans* or how 20 % of X chromosome genes escape inactivation in human females [51, 52]. Once inactivation is established, X chromosome is condensed into facultative heterochromatin and forms a round body at the nuclear periphery [53]. The inactive chromosome possesses repressive chromatin marks and DNA methylation at CpG islands [52, 54, 55].

Genomic imprinting is another epigenetic phenomenon that utilizes lncRNA [56]. Imprinted genes play an important role in mammalian development and therefore their expression has to be tightly regulated [57]. Interestingly, many imprinted gene loci express lncRNAs that play a crucial role in regulating the expression of neighboring imprinted coding genes in *cis* [58]. One such lncRNA involved in genomic imprinting in *cis* is *Air*, which is mono-allelically expressed from the paternal allele. *Air* is known to bind to G9a histone methyltransferase and associate with chromatin to participate on silencing of three imprinted genes: Slc22a3, Slc22a2 and Igf2r. Loss of *Air* leads to bi-allelic expression of Slc22a3 and loss of G9a recruitment to imprinted genes. It has been suggested that Air acts to guide G9a to chromatin at the Slc22a3 promoter [59].

Epigenetic Regulation in *trans*

In contrast, to previous examples, a long intervening lncRNA, HOTAIR, regulates human genes expression in *trans* on a genome-wide scale by associating with chromatin modifying complexes such as polycomb repressive complex (PRC2), LSD1 and CoREST/REST [5, 60–62]. It has been shown that 5′ domain of HOTAIR binds PCR2, whereas a 3′ domain of HOTAIR binds LSD1/CoREST complexes. This way HOTAIR guides PCR2 and LSD1/CoREST to their endogenous targets. Consequently, PRC2 methylates histone H3 lysine 27, whilst LSD1/CoREST demethylates histone H3 at lysine 4. This collectively leads to the loss of active histone marks (H3K4 dimethylation) and the gain of a repressive histone marks (H3K27 trimethylation) at the target loci [62].

Gene Regulation Through lncRNA Transcription

Transcription of lncRNA itself can act as both, a positive (activation) or negative (repression) regulator of gene expression (Fig. 4.5), affecting expression of neighboring genes.

Activation: the act of lncRNA transcription can help to open the chromatin structure of a genetic locus to permit access of transcription machinery to neighboring protein-coding genes. In *fission yeast*, transcription of lncRNAs *UAS1* and *UAS2* have been shown to activate the expression of the *fbp1* gene by this mechanism. Pol II transcribes several species of ncRNAs at the *fbp1* locus during transcriptional activation. The chromatin is progressively converted to an open configuration, which is coupled to translocation of Pol II through the upstream region of the *fbp1* transcriptional start site. It has been shown that transcription through the promoter region is required to make DNA sequence accessible to transcriptional activators and to Pol II [63]. Similar example of gene transcription regulation have been observed within β-globin locus [64].

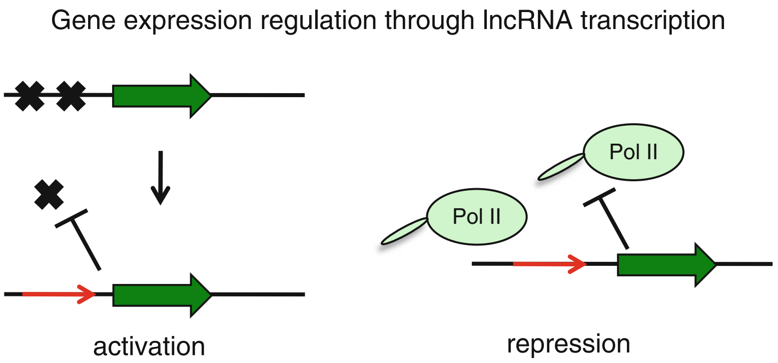


Fig. 4.5 Gene expression regulation through lncRNA transcription. *Block arrows* are protein-coding genes, *red arrows* are lncRNAs. *Block crosses* depict negative transcription factor

Repression: transcription of lncRNAs near protein-coding loci can also act as a negative regulator. The presence of the transcription machinery on the lncRNA gene locus can physically prevent transcription machinery from binding to the protein-coding gene. In *budding yeast*, transcription of the lncRNA *SRG1* inhibits transcription of the overlapping *SER3* gene. This repression occurs by a transcription-interference mechanism in which *SRG1* transcription across the *SER3* promoter interferes with the binding of activators [65]. Such a transcriptional interference process may represent a widespread function for lncRNAs. There seems to be a strong conservation of their promoter regions in contrast to weaker conservation of their transcripts, which is consistent with the act of transcription itself having a greater biological impact than the transcript sequence [7, 66].

Transcriptional Regulation

Protein coding gene expression is tightly regulated process, which involves direct interactions of proteins with other proteins or DNA. Another aspect of the regulation of gene expression comes from an additional layer of complexity consisting of dynamic interactions between RNA, DNA or proteins. Transcription of lncRNAs can regulate the expression of neighboring genes (*in cis* regulation) or can also target distant transcriptional activators or repressors (*in trans* regulation) (Fig. 4.6).

Transcriptional Regulation in *cis*

If lncRNA sequence overlaps through complementarity with the binding site of a transcription factor, the lncRNA transcript can hybridize to this site and so prevent a transcription factor from binding. One such an example is a lncRNA that is transcribed from a minor promoter upstream of the human dihudrofolate reductase

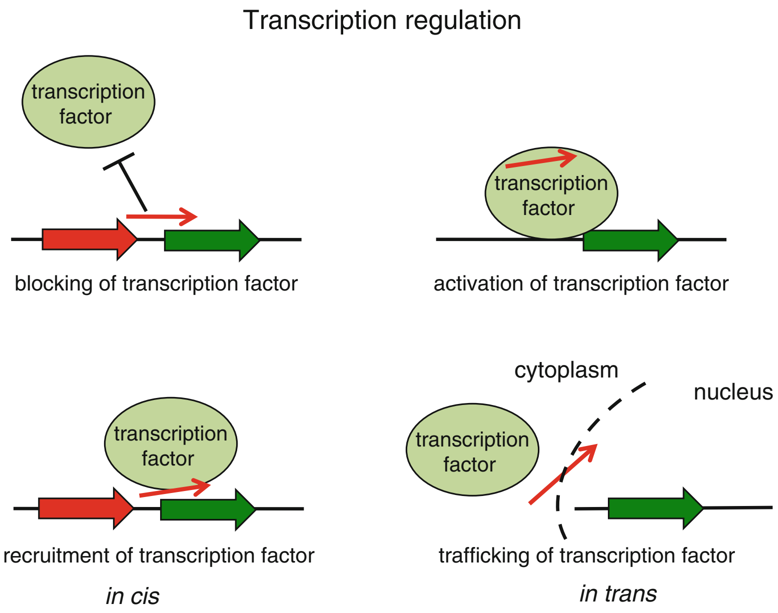

Transcription regulation
transcription factor
blocking of transcription factor
transcription factor
activation of transcription factor
cytoplasm
nucleus
transcription factor
recruitment of transcription factor
transcription factor
trafficking of transcription factor
in cis
in trans

Fig. 4.6 Types of transcriptional regulation by lncRNAs. *Block arrows* are protein-coding genes, *red arrows* are lncRNAs. *Dashed line* depicts nuclear membrane

DHFR gene. Transcription of this full length lncRNA is thought to repress transcription from the major *DHFR* promoter [67] in an RNA dependent manner. The lncRNA binds to the major *DHFR* promoter and the general transcription factor IIB and leads to dissociation of the pre-initiation complex. It is proposed that the single-stranded lncRNA hybridizes to double-stranded DNA in the promoter region to form a triplex structure. Such a structure is predicted to be most concentrated around human promoters [68], but it is unclear whether this is a common mechanism for lncRNA transcriptional repression.

Another type of transcriptional gene regulation by lncRNA is the recruitment of transcription factors. When a lncRNA sequence is located near to transcription factor binding site, the lncRNA transcript may enhance the binding of the transcription factor to promoter region. An example of this type of regulation is the lncRNA called *Evf2*, which regulates two homeodomain genes, *Dlx5* and *Dlx6*, involved in neuronal differentiation, migration and limb pattering [69]. Single-stranded *Evf2* forms a complex with *Dlx2*, another homeodomain protein. This *Evf2-Dlx2* complex activates *Dlx5/6* enhancer, by a yet unknown mechanism.

Transcriptional Regulation in *trans*

Another way that lncRNA may regulate transcription is through their affect on transcription factors trafficking in the cell. In particular, lncRNA can either enhance transcription factor access to DNA binding sites or prevent it, as in the case of the lncRNA *NRON*. This lncRNA prevents the transcription factor *NFAT* (nuclear factor of activated T cells) from entering the nucleus by directly interacting with importin-beta 1, one of the nuclear-cytoplasm transport factors [70]. The *NRON* gene contains three exons and can be alternatively spliced, producing variant transcripts ranging in size from 0.8 to 3.7 kb. Depletion of *NRON* leads to increased levels and activity of *NFAT* in nucleus. Interestingly *NRON*'*s* predicted secondary structure is rich in stem loops, which is conserved between diverse vertebrates and requires further study.

LncRNA can bind to accessory proteins to activate them allosterically, or induce their oligomerization and activation. One such lncRNA is *HSR1* (heat shock RNA-1), which together with an eukaryotic translation-elongation factor 1A, stimulates trimerization of heat-shock factor 1 (*HSF1*) [71]. Trimeric *HSF1* activates heat-shock proteins by binding to their promoters. Formation of *HSR1-HSF1* is induced by heat shock and knockdown of *HSR1* causes cells to become thermo-sensitive. This suggests that HSR1 may be a part of cellular thermo-sensing machinery, resembling a similar mechanism in bacteria.

Post-transcriptional Processing

In addition to all of the above transcriptional mechanisms, many lncRNAs are also involved in post-transcriptional processing of protein-coding mRNAs, including regulation of splicing, editing, transport, translation, and degradation of their corresponding mRNA transcripts.

Natural antisense transcripts (NATs) are typical example of lncRNAs that act to regulate mRNA dynamics. Unlike NATs associated with imprinting genes such as Tsix, Air or HOTAIR, which induce epigenetic changes in chromatin and lead to gene silencing, other NATs can form RNA duplexes to mask key *cis* regulatory elements. This can lead to an alternative splicing pattern of overlapping gene transcripts. For example, the Zeb2/Sip1 NAT is complementary to the 5′ splice site of an intron of the zinc finger Hox mRNA Zeb2, which is involved in epithelial-mesenchymal transition (EMT). Zeb2 NAT is expressed upon EMT and masks the splice site, so blocking spliceosome function. This causes the translation machinery to recognize and bind to an internal ribosome entry site (IRES) in the retained intron resulting in more efficient Zeb2 translation (Fig. 4.7) [72].

More recently, a NAT specific for tyrosine kinase containing immunoglobulin and epidermal growth factor homology domain-1 (Tie-1) was identified in zebrafish, mouse and human. The tie-1 NAT specifically binds tie-1 mRNA in vivo, forming an RNA–RNA duplex. This leads to down-regulation of the Tie-1 protein with consequent specific defects in cellular endothelial contact junctions [73].

In contrast, the expression of beta-secretase-1 (BACE-1) NAT increases stability of BACE-1 mRNA and leads to high production of Abeta (amyloid-beta) 1-42 through a post-transcriptional feed-forward mechanism [74, 75]. In this way BACE-1 NAT acts a positive regulator of Abeta 1-42 through stabilization of BACE-1 mRNA.

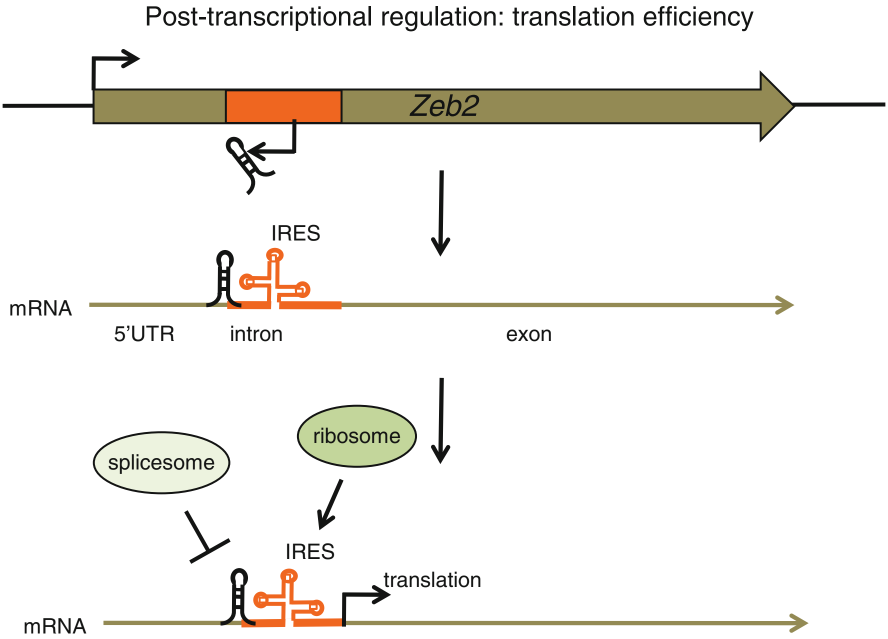


Fig. 4.7 Post-transcriptional regulation by lncRNA. *Brown block arrow* is Zeb2 gene. Intron is colored in *orange*. Hairpins correspond to miRNAs

Alternative splicing is a well-known mechanism of pre-mRNA processing in higher eukaryotes. The serine/arginine (SR) splicing factors regulate cell type specific alternative splicing in a concentration and phosphorylation-dependent manner. How levels of active SR proteins are regulated is not well understood. Recent studies on the long nuclear-retained regulatory RNA (nrRNA) called MALAT (metastasis-associated lung carcinoma transcript 1) implicated its role in alternative splicing [76, 77]. MALAT1 (also known as NEAT2) a 7 kb RNA is localized in nuclear speckles, where it interacts with SR splicing factor, SRSF1 and affects the distribution of other splicing factors. Depletion of MALAT1 changes the alternative splicing profile of multiple endogenous pre-mRNAs. More importantly, MALAT1 regulates the phosphorylation status of SR proteins, thereby regulating pre-mRNA processing via modulation of active SR proteins levels.

Furthermore, there is growing evidence showing that transcripts produced from pseudogenes play an important role in regulating mRNA stability of the gene paralogue. For example, transcripts from the tumor suppressor pseudogene of *PTEN* (*PTENP1*) and oncogenic *KRAS* (*KRASP*) regulate levels of their gene counterparts, *PTEN* and *KRAS* [39, 78]. It is biologically relevant to keep the right dosage of *PTEN* in the cell. A number of miRNA and pseudogene transcripts are also directly involved in *PTEN* dosage regulation at a post-transcriptional level. *PTEN* and *PTENP1* 3′ UTRs are highly conserved. *PTENP1* RNA, which is also referred to as competing endogenous RNA (ceRNA), binds to common miRNA preventing their

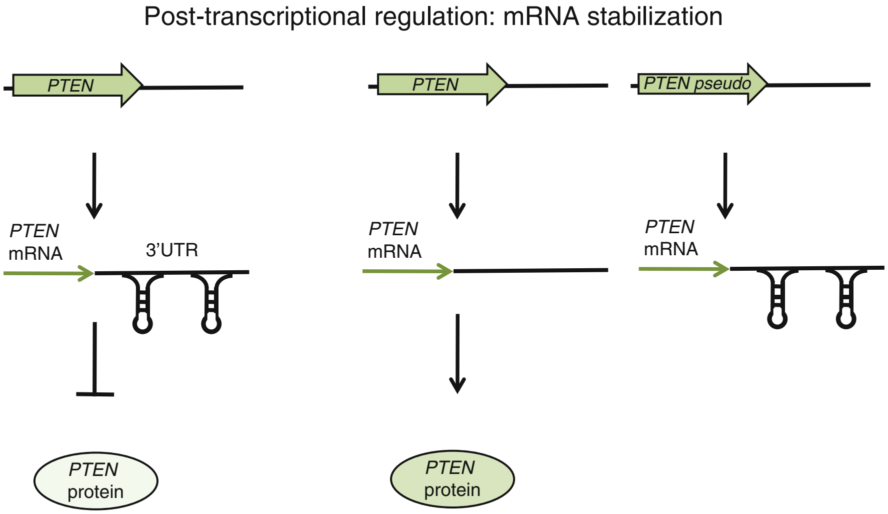


Fig. 4.8 Post-transcriptional regulation by lncRNA. *Block arrows* depict protein coding gene PTEN. Hairpins represent miRNAs

binding to miRNA response element in 3′ UTR of *PTEN*. This competitive binding of *PTENP1* to miRNA results in increased levels of *PTEN* RNA and consequently *PTEN* protein (Fig. 4.8) [40, 79, 80].

Additionally a similar mechanism has been shown for *KRAS* mRNA, which is increased by expression of ceRNAs of *KRASP*. Interestingly, some protein-coding genes, such as *ZEB2*, *VAPA* and *CNOT6L* can also act as ceRNAs.

The initial studies describing links between imprinting and X chromosome inactivation were based on discovery of the *H19* and *Xist* RNAs. The *H19* gene encodes a 2.3 kb lncRNA that is expressed exclusively from the maternal allele and is spliced, polyadenylated and exported into cytoplasm where it accumulates [81]. *H19* cause imprinting of its counterpart protein-coding gene, the insulin-like growth factor 2, *IGF2*. Recent studies revealed that *H19* is host to an exonic miRNA, *miR-675*, which is also imprinted and maternally expressed. *H19* and *miR675* are conserved across mammalian species, suggesting that both are selected [82]. It is proposed that *H19* might act through the nonsense mediated RNA decay pathway. Indeed, a key component of this pathway has been shown to regulate the levels of *H19* RNA during embryonic stem cell differentiation [83].

lncRNA and Disease

Recent and rapid progress in lncRNA research reveals a growing body of evidence that lncRNA play an important role in variety of normal physiological processes. Consequently their mis-regulated expression contributes to numerous diseases, including cancer.

lncRNA and Cancer

New technological approaches, such as genome-wide studies, RIP-RNA sequencing, gene expression screens, region-targeted assays and gene knock-down/knock-out experiments all contribute to the determination of lncRNA function in pathogenesis. Accumulating data show that lncRNAs are indeed involved in carcinogenesis, invasion and metastasis. Based on their function, lncRNAs can be divided into two major categories: oncogenic and tumor suppressor classes.

Oncogenic lncRNAs

Some lncRNAs, referred to as oncogenic transcripts, can regulate cellular pathways that lead to oncogenesis. Recent studies identify more and more of onco-lncRNAs such as *KRASP*, *HULC*, *HOTAIR*, *MALAT1*/*NEAT1*, *p15AS*, *ANRIL*, *H19*, *SRA1*, *p21NAT* or *RICTOR*. Some lncRNAs can act as oncogenic as well as tumor suppressor transcripts, depending on cellular context.

The lncRNA, referred to as cancer metastasis-associated lung adenocarcinoma transcript, *MALAT1*, was identified in non-small-cell lung cancer [84]. *MALAT1* is abundant and plays a key role in cell proliferation, migration and invasion. It localizes predominantly in nuclear speckles in a transcription dependent manner to regulate mRNAs post-transcriptional processing such as alternative splicing [76, 85]. *MALAT1* is also up-regulated in other types of cancer, including breast, prostate, liver and colon [84, 86–90]. Furthermore, higher expression of *MALAT1* is associated with metastatic tumors where it is correlated with poor prognosis. Recent studies demonstrate that *MALAT1* is involved in cell mobility at it targets genes, required for cell migration, in order to regulate their gene expression at both a transcriptional and post-transcriptional level. However, the underlying mechanism of *MALAT1* in tumor metastatic process remains unclear.

A genome-wide study unveiled associations of multiple genetic variants in a large "gene-desert" region of chromosome 8q24 with susceptibility to prostate cancer (PC). Re-sequencing approaches helped to identify a 13 kb long intron-less lncRNA, termed *PRNCR1* (prostate cancer non-coding RNA1) [91]. Depletion of *PRNCR1* attenuated the viability of PC cells and the *trans*-activation activity of the androgen receptor. Therefore, it has been proposed that *PRNCR1* is involved in prostate carcinogenesis through androgen receptor activity. These findings provide a novel insight into understanding the pathogenesis of genetic factors for prostate cancer.

The lincRNA *HOTAIR* is expressed in many posterior and distal sites during evolution and is highly conserved in vertebrates [60]. *HOTAIR* is over-expressed in metastatic breast cancer and correlates with poor prognosis [61]. De-regulation of *HOTAIR* represses the expression of a subset of cell-to-cell interaction promoting genes, including *JAM2*, *PCDH10*, *PCDHB5* and *EPHA1*. Furthermore, the interaction of *HOTAIR* and *PRC2*, which leads to increased H3K27 trimethylation and silencing of metastasis suppressor genes, is responsible for *HOTAIR* mediated tumor cell invasion and subsequent metastasis. Additionally, increased levels of *HOTAIR* were detected in hepatocellular carcinoma (HCC), suggesting that *HOTAIR* could be a candidate biomarker for tumor recurrence prediction. Depletion of *HOTAIR* in liver cancer cells results in reduced cell viability and sensitized apoptosis induced by *TNF-alpha* [92]. These studies indicate that lincRNAs have active roles in modulating the cancer epigenome and could be an important predictor for cancer outcome as well as novel targets for cancer therapy.

Loss of imprinting (LOI) is involved in a number of human hereditary diseases and cancers [93]. Disruption in the expression of imprinted genes such as *H19*, *p57*, *IGF2* and *KvLQT1*, results in almost 80 % of Beckwith–Wiedemann syndrome (BWS). About 5–10 % of BWS patients are predisposed to a number of childhood tumors. *Kcnq1ot1* is an imprinted antisense lncRNA, which is about 60 kb long and possesses a silencing domain at its 5′ end [94]. *Kcnq1ot1* transcript is associated with multiple chromosomal rearrangements in BWS. Its abnormal expression was observed in 50 % of BWS patients and 53 % of colorectal cancers [95–97]. Loss of *Kcnq1ot1* imprinting is accompanied by loss of methylation of the control element, a CpG-island called *KvDMR1*. *KvDMR1* contains the promoter for the paternally expressed *Kcnq1ot1*. Disruption of this promoter abolishes *Kcnq1ot1* transcripts leading to activation of neighboring genes such as tumor repressor *CDKN1C* [98]. These data suggest that abnormal expression of *Kcnq1ot1* contributes to carcinogenesis.

Recent studies report consistent differences in the expression of sense and antisense transcripts between normal and neoplastic cells. A group of genes that generate NATs in normal, but not cancer cells are involved in essential metabolic processes. Altered ratio of sense and antisense transcription contributes to tumorigenesis and cancer progression [99–103]. For example, leukemic cells express higher amounts of antisense p15 NATs and smaller amounts of its partner p15 sense mRNA than normal lymphocytes. Many NAT lncRNAs may have relevance to the cancer genes, including p21, p53, E-cadherin or myc [104]. Thus, it is proposed that tumorigenic NATs are a trigger for heterochromatin formation and DNA methylation in tumor suppressor silencing.

Tumor Suppressor lncRNAs

Some lncRNAs are found to function as tumor suppressors, resembling some protein-coding genes. This group of lncRNAs includes *MEG3*, *GAS5*, *lincRNA-p21*, *PTENP1*, *TERRA*, *CCND1* and *TUG1*.

MEG3 is a lncRNA transcript of a maternally imprinted gene, which is expressed in normal human cells. Loss of *MEG3* was found in meningiomas and adenomas of gonadotroph origins [105, 106]. *MEG3* is a positive regulator of the tumor suppressor gene, p53. Ectopic expression of *MEG3* up-regulates p53 protein levels and dramatically induces p53 transcription. Furthermore, *MEG3* selectively enhances p53 binding to its target promoter, such as *GDF15*. Expression of *MEG3* is able to inhibit cell proliferation in the absence of p53. All these data suggest that lncRNA *MEG3* functions as a tumor suppressor in both a p53-dependent and p53-independent manner [107, 108].

LincRNA-p21 is another example of a tumor suppressor lncRNA, whose expression is directly induced by the p53-signaling pathway. *LincRNA-p21* is required for global repression of genes that interfere with p53 function to regulate cellular apoptosis. This occurs through physical interaction with RNA-binding protein hnRNP-K leading to its localization on gene promoters, which are thought to be repressed in a p53-dependent manner [109].

GAS5 (growth arrest-specific 5) is tumor suppressor lncRNA, which regulates normal growth in lymphocytes. Depletion of *GAS5* inhibits apoptosis and maintains rapid cell cycling, which indicates that its expression is necessary for normal growth arrest. *GAS5* regulates the expression of a critical group of genes with tumor suppressive functions. Additionally, several snoRNAs are transcribed solely from *GAS5* introns. Under starvation, *GAS5* directly interacts with the DNA binding domain of glucocorticoid receptor (GR), leading to inhibition of GR binding to its target gene promoters. Such repression is not limited only to GR, but applies also to other members of the nuclear receptor family. Interestingly, *GAS5* is significantly downregulated in breast cancer cells [110–112].

Cis-acting lncRNA, *CCND*, originates from the promoter of the *CCND1* gene encoding cyclin D1 protein. Upon induction, *CCND* lncRNA transcript is tethered to the *CCND1* promoter and so inhibits *CCND1* expression. Cyclin D1 is frequently over-expressed in human tumors. Therefore, it is proposed that *CCND1* transcript functions as a tumor suppressor to repress tumorigenesis [113].

Expression of the telomere-related lncRNA, *TERRA*, is highly dependent on development, nuclear reprogramming, telomere length, cellular stresses and chromatin structure. Many abnormal telomere phenotypes in aging and cancer cells are linked to mis-regulated expression of *TERRA*. Low levels of *TERRA* have been observed in the tumor-derived and in vitro immortalized cell lines. It has been proposed that *TERRA*-regulated telomere length plays an important role in tumor development [114–117].

lncRNA and Other Diseases

LncRNAs in the context of their cellular function can also be involved in diseases other than cancer.

Patients with *SCA8* have a trinucleotide expansion in an lncRNA called ataxin 8, which is antisense to the *KLHL1* gene. Involvement of this type of mutation in disease progression was confirmed in mouse model transgenic mice with this repeat expansion displaying a progressive neurological phenotype similar to human *SCA8* [118].

An inherited form of alpha-thalassaemia is caused by the translocation of an antisense lncRNA to a neighboring region of the alpha-globin gene (*HBA2*). Induction of this lncRNA results in epigenetic silencing of *HBA2* leading to anemia [119].

The expression of the antisense transcript to *BACE1* gene, as a response to cell stress, leads to progression of the well-studied Alzheimer's disease [74, 120].

Also, psoriasis-associated RNA induced by stress, called *PRINS*, is up-regulated in skin cells of patients with psoriasis. It acts through down-regulation of *G1P3*, gene encoding a protein with anti-apoptotic effect in keratinocytes leading to psoriasis progression [121].

A study using a single-nucleotide polymorphism marker identified a lncRNA called *MIAT* (myocardial infarction-associated transcript) on chromosome 22 in patients with myocardial infarction [122]. Furthermore, genome-wide analysis identified a region encompassing a lncRNA, *ANRIL*, which is linked to coronary artery disease [123, 124].

Overall, identified lncRNAs play a clear role in pathology of various diseases. It remains to be determined what is their specific function and how they are associated with human pathology.

lncRNA as Biomarkers

To date, although our understanding on how lncRNAs cause disease is far from complete, certain features of lncRNAs make them ideal candidates for therapeutic intervention. For example, only a minority of lncRNAs are unstable. LncRNA half-lives vary over a wide range, comparable to that of mRNAs. Combining half-lives with comprehensive lncRNA annotations hundreds of unstable (half-life < 2 h) intergenic, *cis*-antisense, and intronic lncRNAs, as well as lncRNAs showing extreme stability (half-life > 16 h) were identified. Intergenic and *cis*-antisense RNAs are more stable than those derived from introns [125].

LncRNA expression is elevated in several types of cancers, including human prostate cancer, renal cell carcinomas, breast and ovarian cancer, as well as human lung cancer, suggesting that lncRNAs may become a promising biomarker in disease diagnostics. For example, the prostate specific lncRNA, *DD3*, shows higher specificity than serum prostate-specific antigen (PSA), suggesting that is could be developed into highly specific biomarker [126]. HCC-associated lncRNA, *HULC*, is also upregulated in the blood of hepatocarcinomas, implying its potential use in diagnosis of this type of cancer [127]. Expression of *HOX* specific antisense RNA, *HOTAIR*, is increased in breast tumor cells, suggesting that it may become a powerful predictor of patient outcome such as metastasis and death. *SNORD*-host RNA, *Zfas1* is an antisense transcript of the protein-coding gene *Znfx1*. *Zfas1* is highly expressed in mammary glands and it's obviously down-regulated in breast cancer cells, suggesting its potential for diagnosis of breast cancer [128].

LncRNAs-based biomarkers could also be developed for diseases other than cancer. For example, noncoding transcript for beta-secretase-1 (*BACE1*), which regulates *BACE1* mRNA and protein production, is upregulated in Alzheimer's disease and thus could be exploited as a biomarker [129]. ANRIL lncRNA, is expressed in tissues and cells affected by atherosclerosis, which makes it a potential biomarker for coronary artery disease [130].

Overall it is clear that lncRNAs possess a significant potential for development of new approaches in diagnostics and therapy.

lncRNA and Stem Cell Development

Cellular reprogramming demonstrates the plasticity of cell fates. LncRNAs, whose expressions are linked to pluripotency, are direct targets of key transcription factors [131]. One such a lncRNA (*RoR*) modulates cellular reprogramming, which has been identified by loss-of function and gain-of function approaches. This provided first evidence for the critical function of a lncRNA in the derivation of pluripotent stem cells. LncRNAs also help to regulate development by physically interacting with proteins to coordinate gene expression in embryonic stem cells (ESCs) [132]. This is contrary to the dogma that proteins alone are the key regulators of this process. LncRNA determine the fate of ESCs by keeping them in their unspecialized state or by directing them along a pathway to cellular differentiation.

lncRNA and Immunity

Whole-transcriptome analysis has shown that lncRNAs are associated with diverse biological processes in different tissues and are also involved in the host response to viral infection and innate immunity [6]. Also a recent study revealed altered expression of lncRNA during CD8+ T cell differentiation upon antigen recognition [133]. Likewise, eight mRNA-like lncRNAs were differentially expressed in virus-infected birds [134]. Whole-transcriptome analysis of severe acute respiratory syndrome in coronavirus-infected lung samples shows that there is a widespread differential regulation of lncRNAs in response to viral infection [135]. All of this suggests that lncRNAs are involved in regulating the host response in virus-infected cells, including innate immunity.

Concluding Remarks

The discovery of lncRNAs has changed our view of the complexity of the mammalian transcriptome. LncRNAs are becoming widely recognized as key regulators of protein-coding gene expression and so provide an additional layer of transcriptional control. To date, lncRNAs have been shown to be involved in many different stages of gene expression regulation. This diversity in function suggests that lncRNAs will ultimately be found to participate at all levels of transcriptional control, from nuclear localization of transcription factors to transcriptional termination. Several lncRNAs have been implicated in the mediation of chromatin structure. Enabling the accessibility of the genome to Pol II and its associated factors is the most efficient way to activate or repress transcription. LncRNAs also function in X chromosome inactivation and genomic imprinting through chromatin remodeling.

Future discoveries may struggle to identify additional transcriptional regulatory lncRNA that share a function with known lncRNAs, because different RNAs can have similar functions even though they lack detectable sequence similarity.

Real challenges lie in determining the biological significance of lncRNAs-protein interaction. Scientists have to clearly demonstrate biological roles of particular lncRNAs and relate them to their associated transcriptional units. It is peculiar that lncRNAs are not evolutionary conserved, they are expressed in very low levels and their knock-out don't show a clear phenotype. Therefore, their biological significance remains an open topic for a further analysis.

The de-regulation of lncRNA expression in the context of cell pathology represents a new layer of complexity in the molecular architecture of human disease. Several lines of evidence have suggested that even small-scale mutations can affect lncRNA structure and function. Future studies need to elucidate the mechanism by which disease-causing mutations in lncRNA functional motifs can affect its regulatory domains and thereby contribute to disease pathology.

Future research of lncRNA may lead to discoveries of their biological functions and ultimately propose new RNA-based targets for the prevention and treatment of human disease.

Acknowledgement I am grateful to Prof Nick J Proudfoot for constructive comments on this manuscript.

This work was supported by MRC Career Development Award and by L'Oreal/UNESCO Woman in Science UK and Ireland award to Monika Gullerova.

References

1. Ponting CP, Belgard TG (2010) Transcribed dark matter: meaning or myth? Hum Mol Genet 19:R162–R168
2. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B et al (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25:1915–1927
3. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A et al (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46–58
4. Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–641
5. Khalil AM, Guttman M, Huarte M, Garber M, Raj A et al (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106:11667–11672
6. Guttman M, Amit I, Garber M, French C, Lin MF et al (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227
7. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563
8. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155–159
9. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H et al (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573
10. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. Science 322:1855–1857
11. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560

12. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837
13. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J et al (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28:503–510
14. Schorderet P, Duboule D (2011) Structural and functional differences in the long non-coding RNA hotair in mouse and human. PLoS Genet 7:e1002071
15. Wang J, Zhang J, Zheng H, Li J, Liu D et al (2004) Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature 431:1 p (following 757; discussion following 757)
16. Duret L, Chureau C, Samain S, Weissenbach J, Avner P (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312:1653–1655
17. Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB et al (2008) A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PLoS One 3:e2521
18. Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB et al (2007) A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics 8:39
19. Conley AB, Miller WJ, Jordan IK (2008) Human cis natural antisense transcripts initiated by transposable elements. Trends Genet 24:53–56
20. Ni T, Tu K, Wang Z, Song S, Wu H et al (2010) The prevalence and regulation of antisense transcripts in Schizosaccharomyces pombe. PLoS One 5:e15271
21. Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ et al (2011) Comparative functional genomics of the fission yeasts. Science 332:930–936
22. Yassour M, Pfiffner J, Levin JZ, Adiconis X, Gnirke A et al (2010) Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. Genome Biol 11:R87
23. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M et al (2005) Antisense transcription in the mammalian transcriptome. Science 309:1564–1566
24. Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K (2005) Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res 15:463–474
25. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S et al (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154
26. Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res 13: 1324–1334
27. Gingeras TR (2007) Origin of phenotypes: genes and transcripts. Genome Res 17:682–690
28. Wahlestedt C (2006) Natural antisense and noncoding RNA transcripts as potential drug targets. Drug Discov Today 11:503–508
29. Sun M, Hurst LD, Carmichael GG, Chen J (2005) Evidence for a preferential targeting of 3′-UTRs by cis-encoded natural antisense transcripts. Nucleic Acids Res 33:5533–5543
30. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME et al (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell 121:725–737
31. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM et al (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature 457:1038–1042
32. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S et al (2009) Bidirectional promoters generate pervasive transcription in yeast. Nature 457:1033–1037
33. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS et al (2008) RNA exosome depletion reveals transcription upstream of active human promoters. Science 322:1851–1854
34. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB et al (2008) Divergent transcription from active promoters. Science 322:1849–1851

35. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–1848
36. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182–187
37. Orom UA, Shiekhattar R (2011) Long non-coding RNAs and enhancers. Curr Opin Genet Dev 21:194–198
38. Faulkner GJ, Carninci P (2009) Altruistic functions for selfish DNA. Cell Cycle 8:2895–2900
39. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ et al (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465:1033–1038
40. Tay Y, Kats L, Salmena L, Weiss D, Tan SM et al (2011) Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell 147:344–357
41. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A 105:716–721
42. Babak T, Blencowe BJ, Hughes TR (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinformatics 8:33
43. Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. Mol Cell 43:904–914
44. Nagano T, Fraser P (2011) No-nonsense functions for long noncoding RNAs. Cell 145:178–181
45. Clark MB, Mattick JS (2011) Long noncoding RNAs in cell biology. Semin Cell Dev Biol 22:366–376
46. Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF (2009) RNA regulation of epigenetic processes. Bioessays 31:51–59
47. Clemson CM, McNeil JA, Willard HF, Lawrence JB (1996) XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. J Cell Biol 132:259–275
48. Chaumeil J, Le Baccon P, Wutz A, Heard E (2006) A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. Genes Dev 20:2223–2237
49. Jonkers I, Monkhorst K, Rentmeester E, Grootegoed JA, Grosveld F et al (2008) Xist RNA is confined to the nuclear territory of the silenced X chromosome throughout the cell cycle. Mol Cell Biol 28:5583–5594
50. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322:750–756
51. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434:400–404
52. Khalil AM, Driscoll DJ (2007) Trimethylation of histone H3 lysine 4 is an epigenetic mark at regions escaping mammalian X inactivation. Epigenetics 2:114–118
53. Barr ML, Bertram EG (1949) A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. Nature 163:676
54. Panning B (2004) X inactivation in mouse ES cells: histone modifications and FISH. Methods Enzymol 376:419–428
55. Heard E, Rougeulle C, Arnaud D, Avner P, Allis CD et al (2001) Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. Cell 107:727–738
56. Reik W, Murrell A (2000) Genomic imprinting. Silence across the border. Nature 405:408–409
57. Li Y, Sasaki H (2011) Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. Cell Res 21:466–473
58. Mohammad F, Mondal T, Kanduri C (2009) Epigenetics of imprinted long noncoding RNAs. Epigenetics 4:277–286
59. Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC et al (2008) The air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322:1717–1720

60. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X et al (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129:1311–1323
61. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM et al (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464:1071–1076
62. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK et al (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science 329:689–693
63. Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T et al (2008) Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature 456:130–134
64. Gribnau J, Diderich K, Pruzina S, Calzolari R, Fraser P (2000) Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. Mol Cell 5:377–386
65. Martens JA, Laprade L, Winston F (2004) Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature 429:571–574
66. Ponjavic J, Ponting CP (2007) The long and the short of RNA maps. Bioessays 29:1077–1080
67. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445: 666–670
68. Goni JR, de la Cruz X, Orozco M (2004) Triplex-forming oligonucleotide target sequences in the human genome. Nucleic Acids Res 32:354–360
69. Feng J, Bi C, Clark BS, Mady R, Shah P et al (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev 20:1470–1484
70. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG et al (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science 309:1570–1573
71. Shamovsky I, Ivannikov M, Kandel ES, Gershon D, Nudler E (2006) RNA-mediated response to heat shock in mammalian cells. Nature 440:556–560
72. Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB et al (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes Dev 22:756–769
73. Li K, Blum Y, Verma A, Liu Z, Pramanik K et al (2010) A noncoding antisense RNA in tie-1 locus regulates tie-1 function in vivo. Blood 115:133–139
74. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG et al (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med 14:723–730
75. Jiang Y, Mullaney KA, Peterhoff CM, Che S, Schmidt SD et al (2010) Alzheimer's-related endosome dysfunction in down syndrome is Abeta-independent but requires APP and is reversed by BACE-1 inhibition. Proc Natl Acad Sci U S A 107:1630–1635
76. Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T et al (2010) A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. EMBO J 29:3082–3093
77. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q et al (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 39:925–938
78. He L (2010) Posttranscriptional regulation of PTEN dosage by noncoding RNAs. Sci Signal 3:pe39
79. Karreth FA, Tay Y, Perna D, Ala U, Tan SM et al (2011) In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. Cell 147:382–395
80. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O et al (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 147:358–369

81. Bartolomei MS, Zemel S, Tilghman SM (1991) Parental imprinting of the mouse H19 gene. Nature 351:153–155
82. Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D et al (2008) Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. Nat Genet 40:971–976
83. Ciaudo C, Bourdet A, Cohen-Tannoudji M, Dietz HC, Rougeulle C et al (2006) Nuclear mRNA degradation pathway(s) are implicated in Xist regulation and X chromosome inactivation. PLoS Genet 2:e94
84. Ji P, Diederichs S, Wang W, Boing S, Metzger R et al (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22:8031–8041
85. Bond CS, Fox AH (2009) Paraspeckles: nuclear bodies built on long noncoding RNA. J Cell Biol 186:637–644
86. Guffanti A, Iacono M, Pelucchi P, Kim N, Solda G et al (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 10:163
87. Lin R, Maeda S, Liu C, Karin M, Edgington TS (2007) A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. Oncogene 26:851–858
88. Luo JH, Ren B, Keryanov S, Tseng GC, Rao UN et al (2006) Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas. Hepatology 44:1012–1024
89. Schmidt LH, Spieker T, Koschmieder S, Humberg J, Jungen D et al (2011) The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. J Thorac Oncol 6:1984–1992
90. Tano K, Mizuno R, Okada T, Rakwal R, Shibato J et al (2010) MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. FEBS Lett 584:4575–4580
91. Chung S, Nakagawa H, Uemura M, Piao L, Ashikawa K et al (2011) Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. Cancer Sci 102:245–252
92. Yang Z, Zhou L, Wu LM, Lai MC, Xie HY et al (2011) Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. Ann Surg Oncol 18:1243–1250
93. Feinberg AP, Tycko B (2004) The history of cancer epigenetics. Nat Rev Cancer 4:143–153
94. Mohammad F, Pandey RR, Nagano T, Chakalova L, Mondal T et al (2008) Kcnq1ot1/Lit1 noncoding RNA mediates transcriptional silencing by targeting to the perinucleolar region. Mol Cell Biol 28:3713–3728
95. Lee MP, DeBaun MR, Mitsuya K, Galonek HL, Brandenburg S et al (1999) Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. Proc Natl Acad Sci U S A 96:5203–5208
96. Mitsuya K, Meguro M, Lee MP, Katoh M, Schulz TC et al (1999) LIT1, an imprinted antisense RNA in the human KvLQT1 locus identified by screening for differentially expressed transcripts using monochromosomal hybrids. Hum Mol Genet 8:1209–1217
97. Nakano S, Murakami K, Meguro M, Soejima H, Higashimoto K et al (2006) Expression profile of LIT1/KCNQ1OT1 and epigenetic status at the KvDMR1 in colorectal cancers. Cancer Sci 97:1147–1154
98. Horike S, Mitsuya K, Meguro M, Kotobuki N, Kashiwagi A et al (2000) Targeted disruption of the human LIT1 locus defines a putative imprinting control element playing an essential role in Beckwith-Wiedemann syndrome. Hum Mol Genet 9:2075–2083
99. Maruyama R, Shipitsin M, Choudhury S, Wu Z, Protopopov A et al (2012) Altered antisense-to-sense transcript ratios in breast cancer. Proc Natl Acad Sci U S A 109:2820–2824
100. Grigoriadis A, Oliver GR, Tanney A, Kendrick H, Smalley MJ et al (2009) Identification of differentially expressed sense and antisense transcript pairs in breast epithelial tissues. BMC Genomics 10:324

101. Numata K, Osada Y, Okada Y, Saito R, Hiraiwa N et al (2009) Identification of novel endogenous antisense transcripts by DNA microarray analysis targeting complementary strand of annotated genes. BMC Genomics 10:392
102. Kohno K, Chiba M, Murata S, Pak S, Nagai K et al (2010) Identification of natural antisense transcripts involved in human colorectal cancer development. Int J Oncol 37:1425–1432
103. Monti L, Cinquetti R, Guffanti A, Nicassio F, Cremona M et al (2009) In silico prediction and experimental validation of natural antisense transcripts in two cancer-associated regions of human chromosome 6. Int J Oncol 34:1099–1108
104. Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J et al (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature 451:202–206
105. Gejman R, Batista DL, Zhong Y, Zhou Y, Zhang X et al (2008) Selective loss of MEG3 expression and intergenic differentially methylated region hypermethylation in the MEG3/ DLK1 locus in human clinically nonfunctioning pituitary adenomas. J Clin Endocrinol Metab 93:4119–4125
106. Zhang X, Gejman R, Mahta A, Zhong Y, Rice KA et al (2010) Maternally expressed gene 3, an imprinted noncoding RNA gene, is associated with meningioma pathogenesis and progression. Cancer Res 70:2350–2358
107. Benetatos L, Vartholomatos G, Hatzimichael E (2011) MEG3 imprinted gene contribution in tumorigenesis. Int J Cancer 129:773–779
108. Zhou Y, Zhong Y, Wang Y, Zhang X, Batista DL et al (2007) Activation of p53 by MEG3 non-coding RNA. J Biol Chem 282:24731–24742
109. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ et al (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142:409–419
110. Mourtada-Maarabouni M, Pickard MR, Hedge VL, Farzaneh F, Williams GT (2009) GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. Oncogene 28:195–208
111. Mourtada-Maarabouni M, Hedge VL, Kirkham L, Farzaneh F, Williams GT (2008) Growth arrest in human T-cells is controlled by the non-coding RNA growth-arrest-specific transcript 5 (GAS5). J Cell Sci 121:939–946
112. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci Signal 3:ra8
113. Kim JK, Diehl JA (2009) Nuclear cyclin D1: an oncogenic driver in human cancer. J Cell Physiol 220:292–296
114. Feuerhahn S, Iglesias N, Panza A, Porro A, Lingner J (2010) TERRA biogenesis, turnover and implications for function. FEBS Lett 584:3812–3818
115. Caslini C (2010) Transcriptional regulation of telomeric non-coding RNA: implications on telomere biology, replicative senescence and cancer. RNA Biol 7:18–22
116. Luke B, Lingner J (2009) TERRA: telomeric repeat-containing RNA. EMBO J 28: 2503–2510
117. Ng LJ, Cropley JE, Pickett HA, Reddel RR, Suter CM (2009) Telomerase activity is associated with an increase in DNA methylation at the proximal subtelomere and a reduction in telomeric transcription. Nucleic Acids Res 37:1152–1159
118. Moseley ML, Zu T, Ikeda Y, Gao W, Mosemiller AK et al (2006) Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. Nat Genet 38:758–769
119. Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H et al (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat Genet 34:157–165
120. Khalil AM, Faghihi MA, Modarresi F, Brothers SP, Wahlestedt C (2008) A novel RNA transcript with antiapoptotic function is silenced in fragile X syndrome. PLoS One 3:e1486
121. Sonkoly E, Bata-Csorgo Z, Pivarcsi A, Polyanka H, Kenderessy-Szabo A et al (2005) Identification and characterization of a novel, psoriasis susceptibility-related noncoding RNA gene, PRINS. J Biol Chem 280:24159–24167

122. Ishii N, Ozaki K, Sato H, Mizuno H, Saito S et al (2006) Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J Hum Genet 51:1087–1099
123. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R et al (2007) A common allele on chromosome 9 associated with coronary heart disease. Science 316:1488–1491
124. Pasmant E, Laurendeau I, Heron D, Vidaud M, Vidaud D et al (2007) Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res 67:3963–3969
125. Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E et al (2012) Genome-wide analysis of long noncoding RNA stability. Genome Res 22:885–898
126. Hessels D, Klein Gunnewiek JM, van Oort I, Karthaus HF, van Leenders GJ et al (2003) DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. Eur Urol 44:8–15, discussion 15–16
127. Panzitt K, Tschernatsch MM, Guelly C, Moustafa T, Stradner M et al (2007) Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. Gastroenterology 132:330–342
128. Askarian-Amiri ME, Crawford J, French JD, Smart CE, Smith MA et al (2011) SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. RNA 17:878–891
129. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L et al (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32:232–246
130. Broadbent HM, Peden JF, Lorkowski S, Goel A, Ongen H et al (2008) Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. Hum Mol Genet 17:806–814
131. Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K et al (2010) Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet 42:1113–1117
132. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK et al (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477:295–300
133. Pang M, Woodward AW, Agarwal V, Guan X, Ha M et al (2009) Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). Genome Biol 10:R122
134. Ahanda ML, Ruby T, Wittzell H, Bed'Hom B, Chausse AM et al (2009) Non-coding RNAs revealed during identification of genes involved in chicken immune responses. Immunogenetics 61:55–70
135. Peng X, Gralinski L, Armour CD, Ferris MT, Thomas MJ et al (2010) Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. MBio 1:e00206-10

Chapter 5
Unveiling Transposable Elements Function to Enrich Knowledge for Human Physiology and Disease Pathogenesis

Ioannis S. Vizirianakis, Elsa P. Amanatiadou, and Sotirios S. Tezias

Introduction

Transposable elements (TEs) are DNA sequences that have the ability to move from one chromosomal location to another through their integration into the genome at different sites. TEs have been found in virtually all eukaryotic organisms, covering 3–80 % of their genomes and considerably influencing evolutionary history [1, 2]. In humans, TEs have been estimated to occupy approximately 45 % of genomic sequence, a proportion being one of the highest among mammals. However, as it has been proposed, it is possible that this impressive number represents an underestimation, since sequences of ancient TEs may have deteriorated beyond recognition [3].

TEs can be grouped in two major classes [4]. Class II elements or DNA transposons account for roughly 3 % of the genome and move by a cut and paste mechanism through an element encoded transposase. There is no indication of their activity to date which is believed to have subsided 37 million years ago [5, 6]. Class I elements or retrotransposons move by a copy and paste mechanism through an RNA intermediate which is reverse transcribed and then inserted at a new site in the host genome. This process is known as retrotransposition. Retrotransposons can be further categorized in two major subclasses: long terminal repeat retrotransposons (LTR) and non-LTR retrotransposons. LTR retroelements include endogenous retroviruses (ERVs) and consist 8 % of the human genome. Although inactive in humans for millions of years their activity is significant in rodent germline [7]. The LTR retrotransposon subclass has a mode of retrotransposition very similar to retroviruses. Non-LTR retroelements include long interspersed nuclear elements (LINEs), short interspersed elements (SINEs) such as Alu elements and hybrid SINE-R/VNTR

I.S. Vizirianakis, Ph.D. (✉) • E.P. Amanatiadou, M.Sc. • S.S. Tezias, M.Sc.
Laboratory of Pharmacology, Department of Pharmaceutical Sciences,
Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece
e-mail: ivizir@pharm.auth.gr; elza3782@hotmail.com; vixarionas2003@yahoo.co.uk

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_5

(variable number of tandem repeat)/Alu elements (SVAs) which are all evidently capable of retrotransposition in the human genome. LINEs are termed autonomous, since they encode the proteins required for their mobilisation. All the other elements are non-autonomous, since they encode no proteins and parasitize LINE machinery in order to achieve their genome mobilization.

The LINE-1 (L1) family is represented by more than 500,000 copies in the human genome comprising an estimated 17 % [8]. Most of these copies are inactive due to accumulated mutations and only 80–100 elements are functional and capable of retrotransposition in human [9]. In contrast, several thousand L1 elements are still active in mice. A full length L1 is 6 kb long and includes a 5′-untranslated region (5′-UTR), two open reading frames (ORF1 and ORF2), a 3′-UTR and a poly(A) tail. The 5′-UTR serves as a RNA-polymerase II promoter while the polyA signal in the 3′-UTR is weak and often read through during transcription allowing 3′-flanking sequences to be co-transcribed along L1 sequence. ORF1 and ORF2 encode proteins with RNA-chaperone and endonuclease/ reverse transcriptase activity respectively. Both of the proteins are required for retrotransposition which takes place through a mechanism called “target-site primed reverse transcription” (TPRT). The synthesized RNA transcript from L1 element transcription is first exported to the cytoplasm in order that ORF1 and ORF2 are produced and then re-enters the nucleus and gets reverse-transcribed directly at the site of integration [10]. ORF1 and ORF2 exhibit a *cis* preference to the RNA they were translated from but also act *in trans* to promote the mobilisation of the non-autonomous retroelements such as Alus [11, 12] and SVAs [13, 14].

More than 1,000,000 copies of Alu elements have been identified so far, a fact that renders them the most abundant repeats in the human genome occupying an estimated 11 %. Alu elements are derived from 7SL RNA, a functional component of the signal recognition particle. B1 SINE elements in mice most closely resemble Alus and also derived from 7SL RNA. Alus are 300 bp long and transcribed by RNA polymerase III. They end in a polyA tail of variable length which has been proved to be critical for retrotransposition as is also the integrity of the internal RNA polymerase III promoter [15]. In addition, it has been demonstrated that Alus require only ORF2 of L1 for their mobilization since expression of L1 elements with mutant ORF1 in ex vivo assays allows Alu mobilization [12]. Alu elements exhibit the highest retrotransposition rate per live births (1/21) among the currently mobilizing elements in humans [16]. Interestingly, Alu sequences are highly polymorphic with respect to their presence or absence among individuals.

SVA elements are intermediate in size (~2 kb) and much fewer than the other retrotransposons such as L1 and Alu (2700 copies, 0.2 % of human genome), since they are hominid-specific and originated less than 25 million years ago [17]. SVAs have a composite structure and are generally believed to be transcribed by RNA polymerase II. They include a hexamer repeat region, an Alu like region, a variable number of tandem repeat (VNTR) region, a short interspersed element of retroviral origin (SINE-R) and a polyA tail. Their mode of mobilization relies on the L1 enzymatic machinery as is also the case with Alu elements [18].

Transposable Elements: Innate Role and Function Controversy

Mobile elements and their impressive prevalence in genomes across species has been the subject of intense scientific debate since their discovery by McClintock [19]. An initially underestimated discovery that was faced with skepticism waited for years for validation and finally led to the appreciation of its significance and a Nobel Prize Award. During this course, various experimental efforts have been successfully carried out, while a lot of light has been shed on the structure, distribution and mode of activity of TEs leaving though their true reason of existence still elusive. Barbara McClintock, whose maize breeding experiments provided the first detailed descriptions of TEs, proposed that the genome is a dynamic entity subject to alteration and rearrangement and also referred to TEs as controlling units suggesting they may serve important regulatory roles.

A lot of efforts describing TE functional characteristics have tried to attribute a central more specific role to their presence and mobilization, ranging from control of embryonic development to the well described contribution in producing genetic variability and promoting the evolutionary process. In certain species such as insects, TEs play a role in fundamental cellular processes like the maintenance and structural integrity of DNA during cell division. In fruit fly and silk worm, a type of transposon is suggested to move bits of DNA to the end of the chromosomes to prevent the loss of telomeres following chromosomal replication [20, 21]. Since TE structure, distribution and mode of activity seems to be variable across species their presence might be accompanied by discrete and differential key roles in each case. Most certainly, and based on knowledge accumulated thus far, the term "junk DNA" is presently at least not representative.

The increasing availability of vast amount of genomic information from multiple species combined with the advancements happening in computational techniques for comparative studies has greatly promoted research concerning the impact of TEs on the evolutionary process. Several examples demonstrate clearly that TEs have served as an important creative force in the evolution of genomes crucially affecting variation of genome size, composition and structure among species. Indeed, it is TE presence that defines the differences in genome size of higher organisms compared to prokaryotes, although the number of existed gene structures increases also [1]. Interestingly, even among mammals there is a correlation of genome size with the TE genomic content. Moreover, the types and families of TEs in vertebrate genomes differ greatly, defined in part by the rate of activity and elimination of ancestral TEs, while inter-individual variation within the same species is also present [5]. Such differences can be actually used as markers in conducting phylogenetic analysis. Apart from the structural genetic diversity that results from TE activity, the impact of the new insertion events is considerable, since genome functionality can be affected in various ways, further driving biodiversity and genome evolution [4].

Mechanisms Through Which TE Activity Affects Gene Expression and Function

The most obvious way in which a retrotransposition event can affect the host genome integrity is insertional mutagenesis. TEs can be inserted into or near genes remodeling the local genomic region and can do so often accompanied by 3′ transduction phenomena. In this case, DNA sequences downstream of a L1 element for example, can be transcribed by RNA polymerase II and co-mobilize due to the presence of a weak L1 termination and polyA signal [22]. The rate at which adjacent genomic regions can be transduced along L1 insertion has been computationally estimated at one every five L1 retrotransposition events [23].

The impact of the insertion event can be detrimental to the host, regardless of whether it involves integration into intron or exon sequences of a gene. Insertion in exon regions can cause the creation of new chimeric mRNA and proteins providing that the integration results in a tolerable phenotype. Insertion in the intron region of genes has been reported to result in mis-splicing, Alu exonization or exon skipping, as well as in reduced mRNA levels due to inefficiency of the RNA polymerase II to transcribe through the transposable element [24, 25]. Another phenomenon observed and associated to the insertion of a mobile element is the deletion of an adjacent genomic sequence. Insertion-mediated deletions caused by Alu and L1 have been confirmed by comparative genomic studies and two mechanisms, TPRT-dependent insertion mediated deletion and endonuclease-independent insertion have been identified that can mediate such deletions of host DNA [26].

In addition, instead of disrupting gene function insertions may alter the expression pattern of nearby genes interfering this way with gene regulation. It has been estimated that TE-derived sequences are contained in the coding region of 4 % of human genes and in 25 % of human promoters [27]. The introduction of functional both splice and polyA sites which can lead to aberrant processing of RNA transcripts, or the introduction of promoter sequences and regulatory regions can lead to changes of gene expression profiles with a potential impact on the host behavior [28–31]. Interestingly, it has been shown that 7–10 % of transcription factor binding sites that have been experimentally characterized derive from repetitive sequences including simple sequence repeats and TEs [32]. The first report of interference with transcriptional control of a gene involves a L1 element residing in the apolipoprotein(a) transcriptional control region (ACR). It was demonstrated in vitro that the L1 element when linked in either orientation to the apolipoprotein(a) minimal promoter can confer a tenfold increase in transcriptional activity [33]. On the other hand, an Alu element at the distal part of the human BRCA2 promoter contains a 221 base-pair silencer region. This region has been found to negatively regulate BRCA2 gene expression in breast cell lines in a tissue specific manner [34].

Genome integrity is directly compromised by the induction of double-strand breaks (DSBs) which are related to L1 ORF2 endonuclease activity. Expression of a transiently transfected L1 element in HeLa cells led to a at least tenfold greater induction of DSBs than the rate of L1 integration under the same transfection conditions and the damage was confirmed to be specific to the L1 endonuclease activity [35]. It is difficult to assess the significance this process might hold in vivo since L1 induced DSBs cannot be distinguished by the ones that arise through different mechanisms. On the other hand, inaccurate repair of DSB DNA lesions can lead to random mutations that further compromise genomic integrity [36, 37].

TEs can also affect host genome long after the integration process is completed. Genetic instability can result through non-allelic homologous recombination (NAHR), which it consists the main disease-causing mechanism for Alu elements. In fact, these post-insertional rearrangements pose a far greater risk to the integrity of the genome than the initial insertion event. Homologous recombination is a fundamental biological mechanism and is highly conserved along species. Normally, programmed homologous recombination occurs once during meiosis. However, the presence of abundant and highly homologous sequences in relatively close proximity, such as Alu elements, increases the potential for mutagenic NAHR. This process can result in genomic deletions, duplications, chromosomal inversions, interchromosomal translocations and NAHR events have been implicated in a number of human genetic diseases as they occur at appreciable rates [26, 38].

Another way in which TE activity can perturb normal gene regulation is by altering the epigenetic state at the site of integration. The recruitment of heterochromatin inducing factors that promote DNA methylation and subsequent histone deacetylation may influence the expression of genes that lie in the vicinity of the integration site, apart from suppressing TE element transcription. It has been reported that in humans, Alu elements are methylated in a more focused manner than LTR and L1 elements which provoke larger chromatin modifying effects [38, 39]. In accordance to the above, it has been suggested that L1 elements may play a role in X-inactivation in mammals. The strong presence of L1 elements in the X-chromosome has been proposed to lead to silencing of the intervening sequences through heterochromatization. Therefore LINE elements may act as modifiers of the epigenetic state of the mammalian genome [40].

Mechanisms Involved in the Regulation of TE Activity Within the Cells

TE mobilization has been clearly demonstrated to affect genome stability and function. Under this perspective, it is not surprising that the hosts have evolved several different mechanisms to limit and regulate TE activity (Fig. 5.1). In particular, it has been suggested through comprehensive analysis of the rate of expansion of L1 families that the most advanced mammalian species have the highest ability to restrict L1 expression [41]. These suppressive mechanisms can act on different stages of the TE amplification cycle, both on the transcriptional, as well as on post-transcriptional level. Of major importance, epigenetic regulation and the RNA interference (RNAi) pathway have also been adopted to limit the negative effects of retrotransposition.

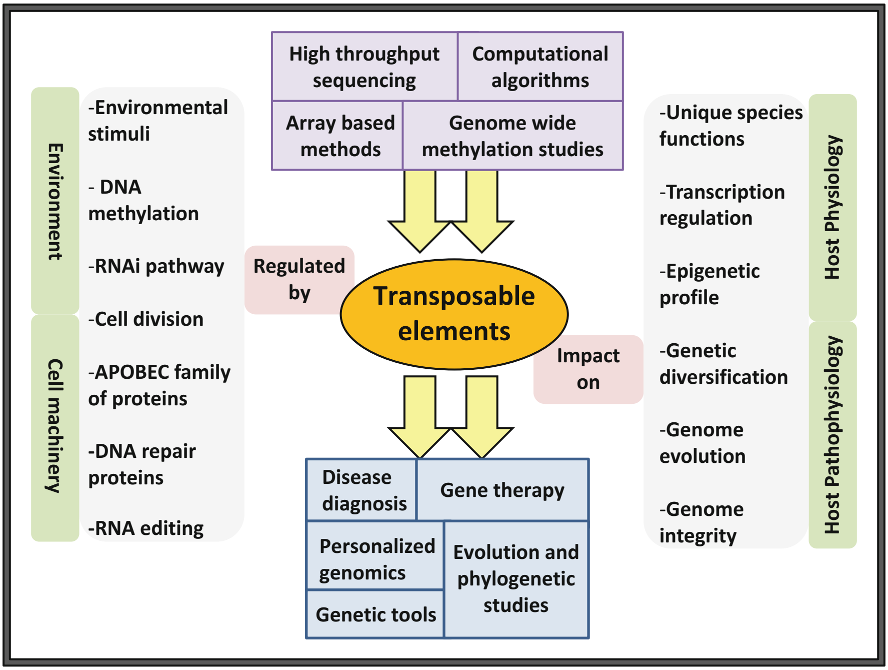


Fig. 5.1 Depiction of principal mechanisms that regulate transposable element activity and the variable ways it can affect host physiology and pathophysiology. In addition, current advancements in the field of DNA sequencing and computational analysis are expected to promote the exploitation of knowledge accumulated and lead to positive outcomes. Many of the concepts presented here are further elaborated on inside the text. *APOBEC* Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like; *RNAi* RNA interference pathway

DNA methylation is the best established mechanism that may influence heterochromatin formation. In mammals, DNA methylation occurs in cytosines embedded in CpG islands. TE promoters are normally transcriptionally-silent due to increased DNA methylation patterns and formation of repressive chromatin states. Actually, methylation measurements of L1 promoters indicate that they range from 20 to 100 % methylated [42]. Most Alus are highly methylated and consequently repressed in differentiated cells (but not in male germ-line cells) [43, 44]. In addition, SVA elements are found in extensively methylated states in most human tissues [45]. SVAs contain numerous CpGs in the VNTR domain and have been suggested to act as species-specific CpG islands [46]. It is noteworthy, that DNA methylation has been proposed to have evolved for the specific purpose of suppressing TE activity [47]. Apart from immediate transcriptional repression, DNA methylation can result in permanent inactivation by C→T deamination [48].

Mutations in genes that belong to the DNA methylation machinery invariably enhance TE transcription and result in reduced viability and fertility. Most representatively, mouse embryos that lack Dnmt1, a DNA methyltransferase, reactivate intracisternal A particles (IAPs), endogenous retroviral sequences representing an important class of TEs, and die before mid-gestation [49, 50]. Dnmt3L knock-out mice exhibit loss of methylation of L1 and LTR elements and a corresponding increase in the levels of expression of their RNA transcripts resulting in failure of the male germline [51]. The same situation exists for proteins that indirectly contribute to the execution of the DNA methylation pathway. Mutation of Lsh, a chromatin remodeling protein that makes DNA accessible to Dnmts, leads to hypomethylation and increase in transcription of IAPs as well as postnatal lethality in mouse embryonic cells [52].

Demethylation of CpG rich promoters is expected to result in enhancement of their activity and increased expression of the TEs. This holds true for several cancer cell lines where the hypomethylation status in these tumor cells has been correlated to increased L1 transcription [53, 54]. This situation is also depicted in the case of chronic myeloid leukemia (CML) where the hypomethylation of L1 elements has been implicated with disease progression [55]. One should bear in mind that demethylation of potent TE promoters has possible implications beyond the activation of retrotransposition. This way, transcription factor balance may be disturbed or transcriptional activation may expand to include genes in the vicinity of demethylated promoters [56]. Activation of the antisense promoter that is located in the 5′ UTR of L1 elements is another possible concern. Interestingly, treatment with azacytidine and decitabine, two hypomethylating agents, has been shown to alter the expression of c-Met oncogene in colon carcinoma and myeloid leukemia cells. The cause of this effect is demethylation of an antisense promoter located in a L1 element in the second intron of the c-Met gene [57]. Histone acetylation on the other hand and its impact on suppressing TE activity is less straightforward and understood so far. It has been reported though that in some cell types after the insertion of engineered L1s the surrounding chromatin is rapidly deacetylated [58].

Of equal importance, the RNAi pathway is presented as a powerful mechanism to control TE activity. RNAi involves small non-coding RNAs of different classes that guide the RNA-induced silencing complex to degrade target transcripts through homology based recognition. The effector complex is made-up of PIWI/Argonaute proteins with RNase H-like activity that can slice single-stranded nucleic acids. The PIWI/Argonaute protein family is subdivided into AGO (Argonaute) families which can bind to small interfering RNAs (siRNAs), microRNAs (miRNAs) in many tissues and PIWI families which can bind to the germ cell specific Piwi-interacting RNAs (piRNAs). Along with DNA methylation, these discrete mechanisms seem to overlap instead of only acting independently in favor of the host genome defense. It has been proposed that miRNAs may direct methylation at the L1 promoter this way maintaining L1 DNA methylation [59]. In addition, mutant male mice for either Mili or Miwi2, two of the PIWI proteins, display loss of DNA methylation at retrotransposon loci and an increase in retrotransposon mRNA expression leading to small testes and sterility [60, 61]. The piRNA pathway appears to hold a dual role in male germline restricting TE activity both by post-transcriptional degradation and by transcriptional silencing through DNA methylation [48, 61, 62].

The female germline differs greatly in that Mili or Miwi2 mutations, even when combined, do not lead to activation of TEs and remain fertile [48]. Alternative mechanisms including the siRNA pathway control TE expression in this case. Although siRNAs were thought to arise primarily from exogenous sources it seems that endogenously produced siRNAs (endo-siRNAs) play a role in restraining TE mobilization. Profiling of the total small RNA population in mouse oocytes uncovered a broad class of endo-siRNAs that was previously unrecognized. This class derived from retroelements including LINE, SINE and LTR retrotransposons and it was suggested that retrotransposons are suppressed through the RNAi pathway in mouse oocytes [63]. siRNAs containing L1, IAP and Mouse Transcript (MT) sequences are produced in oocytes though it has been proposed that they mainly target genes bearing TE repeats in their 3′ UTR region instead of TEs themselves [64]. The control of transposon activity by endo-siRNAs is a well characterized mechanism in *Drosophila* both in somatic tissues and in the germline [65].

The host response and capacity of regulating TE activity is of course multifactorial and reflects the need for accurate tuning and preservation of a critical balance. Epigenetic regulation as well as the RNAi pathway emerged as valuable controlling mechanisms to this end. Several other mechanisms also facilitate the process of TE restraining. To mention a few RNA editing, DNA repair proteins and the APOBEC family of proteins have been shown to modulate TE activity (reviewed in [66]).

Transposable Element Functions in Physiology and Human Pathophysiology

The fact that TE activity can have implications on human disease has been broadly demonstrated and viewed with great interest [24, 26, 67–69]. This way, the ultimate effect of a new transposition event can be neutral, beneficial or harmful to the host, with the propensity of the last to be more easily detected and further examined. Despite the noted effect on creating genetic instability and contributing to human disease pathogenesis, there are numerous examples described in which TEs have had a positive impact on the host physiology resulting occasionally in a favorable phenotype. As an example, a L1-induced transposition event of the cyclophylin gene into the TRIM5 gene is the causative reason owl-monkeys exhibit resistance to human immunodeficiency virus (HIV) 1 infection [70]. Additionally, an Alu insertion polymorphism gene has been associated with protection from dry/atrophic form of age-related macular degeneration [71]. In certain cases host genomes have managed to even exploit TE-encoded proteins in order to support useful functions. This seems to be the case with RAG-1 protein, which is involved in the V(D)J recombination process during antibody production and probably emerged from a DNA transposase [72].

On the other hand, several intricate ways in which TE activity has affected human pathophysiology have been also well documented. The list of human diseases attributed to transposable element mobilization is ever increasing since the initial discovery of a L1 insertion in exon 14 of the factor VIII gene in two unrelated patients with hemophilia A [73]. Approximately a hundred retroelements insertions that cause a wide range of diseases have been recorded accounting for 1/250 (0.4 %) of all disease-causing mutations in human [69]. Numerous examples of all currently mobilizing elements in the human genome (L1, Alu and SVA) exist where de novo insertion led to manifestation of disease. L1 insertions have been reported to result in β-thalassemia, colon cancer and most recently to neurofibromatosis (NF) type I disorder (also known as von Recklinghausen disease) [74]. Alu insertions have in many cases disrupted genes residing on the X chromosome resulting this way in a wide range of human disease (see review article [8]). Interestingly, L1, Alu and SVA elements exhibit a different retrotransposition rate per live births (1/212 for L1, 1/21 for Alu and 1/916 for SVA) that so far coincides with the number of disease-causing mutations attributed to each type of TE (25 for L1, 60 for Alu and 7 for SVA) [8, 16]. In general, the calculated number of retrotransposition events causing monogenic (single-gene involvement) disease has been recently estimated at 500/year [8]. Advancements in whole-genome scale approaches such as high throughput sequencing or novel microarray-based methods are expected to contribute to better accuracy in measurements and thus more insight in the investigation of TE role in human structural variation and disease pathogenesis at the molecular level.

Exploitation of TE Activity for Therapeutic Applications

As an interesting twist to their well-established role in human pathogenesis, TEs have emerged as useful tools for insertional mutagenesis, germline *trans*-genesis, functional genomics applications, and, most importantly, for gene therapy efforts (reviewed in [66]) (Fig. 5.1). In fact, the latest generation in transposon technology has been stratified in a stressful attempt to develop non-viral gene delivery approaches, thus circumventing the problems faced with traditional viral vectors and overcoming the weaknesses that most non-viral vectors exhibit when it comes to gene transduction experimentation studies.

Effective gene therapy involves the process of transducing the gene into the target cells efficiently, succeeding long term expression and also avoiding possible secondary effects such as immune reactions or transformed cell growth. Insertional mutagenesis resulting from the use of integrating viral vectors is a major setback and most non-viral vectors are unable to achieve high and stable expression of the therapeutic gene, highlighting the need for safe and efficient alternatives [75].

DNA transposons replicate through a “cut and paste” mechanism getting excised from a locus and subsequently integrated into another by the transposase protein they encode. The transposase can also act *in trans* on practically any DNA fragment that is flanked by the terminal repeat sequences present at each end of the transposon. This really attractive property led to the development of alternative gene delivery systems in which the DNA sequence of interest can be placed between the transposon terminal repeats and the transposase is supplied in the form of an expression plasmid or mRNA synthesized in vitro.

The fact that DNA transposons are inactive in vertebrates was overcome by genetic engineering in order that these elements obtain the capacity to transpose in mammalian tissues. The transposons Sleeping Beauty (SB), reconstructed from a Tc1/*mariner*-type element, and Piggyback (PB) originating from the cabbage looper *Trichoplusia ni*, demonstrate transposition activity in a wide variety of vertebrate cell lines and species including humans [76]. In a further effort to promote human gene therapy applications, SB100X, a novel hyperactive SB transposase was recently developed through mutations and modifications of the originally reconstructed SB enzyme. The result was a up to 100 times enhancement of its transpositional activity in HeLa cells [77]. In fact, the efficiency in *trans*-gene delivery reaches those of viral vectors and this robustness is extremely useful in applications such as the transfection of primary and other hard to transfect cell types. The SB100X system was able to support 35–50 % stable gene transfer in $CD34^+$ cells enriched in hematopoietic stem or progenitor cells. The gene-marked $CD34^+$ cell population was further transplanted to immunodeficient mice and resulted in long-term engraftment and hematopoietic reconstitution [77].

The interest in applying TE-based systems for effective gene delivery is constantly increasing. These efforts are trying to approach the treatment of several conditions as is junctional epidermolysis bullosa [78], type 1 tyrosinemia [79], hemophilia A and B [80, 81]. Most importantly, the “DNA Advisory Committee” approved in 2008 the first human gene therapy clinical trial that is based on the use of transposons [82]. The aim is to genetically modify T cells with the use of SB transposons in an attempt to treat patients with $CD19^+$ B-lymphoid malignancies. T cells will be co-transfected with a transposon encoding a chimeric antigen receptor (CAR) to enable T cells to recognize lineage-specific tumor antigen, such as CD19, along with a construct expressing an early generation hyperactive SB transposase. The outcome and promise of this effort is eagerly anticipated.

Erythroid Maturation Program of MEL Cells: Lessons Learned from the Blockade of Differentiation by Methylation Inhibitors and the Activation of TEs

Previous work from our laboratory has indicated that the induction of haemoglobin synthesis and terminal erythroid maturation of murine erythroleukemia (MEL) cells in vitro is associated with changes in methylation of RNA species, as well as with alterations in the intracellular concentration of intermediates involved in the active methylation cycle [83–85]. Such a conclusion has gained further support by the fact that N^6-methyladenosine (N^6mAdo) has inhibited commitment of MEL cells to terminal maturation through its intracellular conversion into S-N^6-methyladenosylhomocysteine (N^6-SAH), an active intermediate that affects methylation of RNA and DNA [86]. It has been interesting then to observe that MEL cells exposed to chemical inducers and simultaneously treated with methylation inhibitors (DNA/RNA methylation blockers) although they express the β^{major}

globin gene, however, they: (a) exhibit a blockade to terminal erythroid maturation and (b) produce relatively short polyA$^-$ (non-polyA tail) RNA transcripts accumulated in the cytoplasm that share significant structural homology with 3′-end downstream β^{major} globin gene DNA sequences including the B1 retrotransposon element located within this region (designated as B1-559; see Fig. 5.2) [85, 87, 88]. It is noteworthy to emphasize that B1 repeat element has been found to be located in a DNA region where several consensus binding sequences exist for the erythroid-specific transcription factors GATA-1, AP1/NF-E2 and EKLF. Recently moving toward elucidating the potential functionality of B1-559, we have published data showing that this locus has been: (1) a candidate region encoding the core structure of some of the cloned non-polyA tail RNA species accumulated mainly in the cytoplasm of MEL cells treated with methylation inhibitors; and (2) capable to recruit transcription factors and drive the expression of luciferase gene in cooperation with DNase I hypersensitive site 2 (HS2) derived from locus control region (LCR) of human β-globin gene cluster [89, 90]. Such a direction has been further supported by earlier reports indicated the involvement of Alu-like repetitive sequences in the coordinated post-transcriptional control of gene expression [91], while non-globin DNA sequences located in large distance from the Gγ-globin and other globin genes were found to be potentially homologous to an RNA polymerase III template [92].

In our case, the fact that the B1-559 DNA fragment exhibited potential promoter-like activity in cooperation with HS2 tend to suggest that the identified three consensus sequences for GATA-1, AP1/NF-E2 and EKLF located within this fragment (see Fig. 5.2a) may contribute to transcriptional activation mechanisms. It has been proposed, based on these data, that B1-559 exerts a potential transcriptional role in this part of the genome and thus may affect the expression pattern of β^{major} gene during development, as shown in Fig. 5.2b, by cooperating with enhancer sequences previously known to serve major roles [93–95]. To this end, recent findings implicate the involvement of B1 repeat family in complex regulatory elements in the control of gene expression in the mouse genome. In particular, insertion of a B1 in the 3′-untranslated region of the rabbit β-globin gene generated a construct that conferred transcriptional regulation of this recombinant gene upon its transfection into the cells [91]. Such transcriptional activation of B1-559 also coincides with data implicating B1 retrotransposon elements in the recruitment of transcription factors like PAX6 at discrete binding sites throughout the mouse genome [96]. Moreover, a novel abundant NF-κB-binding site residing in specialized Alu-repetitive elements having the potential for long range transcription regulation has been recently identified by genome analysis [97]. Similar observations were also shown for L1s. P53 DNA binding sites have been detected within L1 which are involved in the regulation of their expression in the mouse genome [98]. Indeed, the transcriptional activation of L1 elements appears to be closely related with hypomethylation of their promoter region, whereas such activity of L1 has been recently shown to contribute to the progression and clinical behavior of chronic myeloid leukemia (CML) [55]. Overall, such functions of TEs to regulate gene activation or silencing represent processes of great interest that are under scrutiny [55, 99]. To this regard, an additional diverse role of TEs in the mouse and human genome was

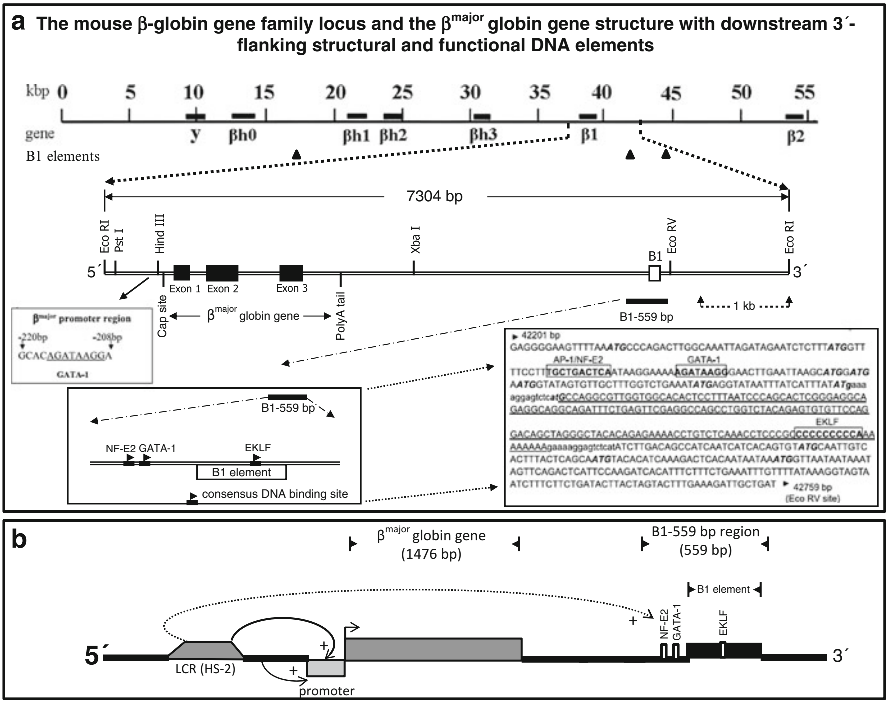

a The mouse β-globin gene family locus and the βmajor globin gene structure with downstream 3´-flanking structural and functional DNA elements
kbp 0 5 10 15 20 25 30 35 40 45 50 55
gene y βh0 βh1 βh2 βh3 β1 β2
B1 elements
7304 bp
Eco RI
Pst I
Hind III
Xba I
Eco RV
Eco RI
B1
5´
3´
Exon 1
Exon 2
Exon 3
Cap site
βmajor globin gene
PolyA tail
B1-559 bp
1 kb
βmajor promoter region
-220bp -208bp
GCACAGATAAGGA
GATA-1
B1-559 bp
NF-E2 GATA-1
EKLF
B1 element
consensus DNA binding site
42201 bp
GAGGGGAAGTTTTAAATGCCCAGACTTGGCAAATTAGATAGAATCTCTTTATGGTT
AP-1/NF-E2
GATA-1
TTCCTTTGCTGACTCAATAAGGAAAAAGATAAGGGAACTTGAATTAAGCATGGATG
AATGGTATAGTGTTGCTTTGGTCTGAAATATGAGGTATAATTTATCATTTATATgaaa
aggagtctcatGCCAGGCGTTGGTGGCACACTCCTTTAATCCCAGCACTCGGGAGGCA
GAGGCAGGCAGATTTCTGAGTTCGAGGCCAGCCTGGTCTACAGAGTGTGTTCCAG
EKLF
GACAGCTAGGGCTACACAGAGAAACCTGTCTCAAACCTCCCGCCCCCCCCCCAAA
AAAAAAgaaaaggagtctcatATCTTGACAGCCATCAATCATCACAGTGTATGCAATTGTC
ACTTTACTCAGCAATGTACACATCAAAGACTCACAATAATAAATGGTTAATAATAAAT
AGTTCAGACTCATTCCAAGATCACATTTCTTTCTGAAATTTGTTTTATAAAGGTAGTA
ATCTTTCTTCTGATACTTACTAGTACTTTGAAAGATTGCTGAT
42759 bp
(Eco RV site)
b
βmajor globin gene
(1476 bp)
B1-559 bp region
(559 bp)
B1 element
+
NF-E2
GATA-1
EKLF
5´
3´
+
+
LCR (HS-2)
promoter

Fig. 5.2 Diagrammatic outline of mouse β-globin gene family locus along with unique structural and functional DNA sequences located within the 3′-flanking sequences of β^{major} globin gene with potential cooperativity to distal enhancer elements to affect transcription. *Panel A*: The depiction of the DNA locus of mouse β-globin gene family is shown at the top along with the B1 repetitive elements identified (shown as *solid triangles*). Moreover, the structural analysis of sequences located at the 3′-flanking region of β^{major} globin gene is indicated with the presence of several consensus sequences for the binding of the NF-E2, GATA-1 and EKLF transcription factors. In addition, this area has several ATG initiation codons (shown in *bold* and *italics*) and a B1 element whose sequence is *underlined*. Finally, the GATA-1 consensus sequence known to exist within the promoter region of β^{major} globin is also illustrated. For more details see [87, 108]. *Panel B*: The influence of close and distant DNA elements on the expression of β^{major} globin gene is presented where besides the known positive cooperative effect of LCR sequences with promoter sequences is well known, it is proposed also a cooperation between the B1 element containing region (B1-559), with LCR to affect β^{major} globin gene transcription based on recent data published from our laboratory shown in [89]. For more details see [89, 90]. (This figure is based on data published from our laboratory and others as shown in [87, 89, 90, 108])

further outlined. In particular, it has been documented that the presence of Alu elements within the 3′-UTR region of genes may lead to their targeted silencing [100]. Furthermore, the creation of two isoforms of rodent natural killer (NK) cell-activating receptor NKG2D gene seems to be driven by a B1 retrotransposon insertion leading to gene regulatory change and its final functional diversification [101]. And consistent with these results, novel transcripts of the human neuronal apoptosis inhibitory protein (NAIP) gene has been shown to arise from transcription start sites that initiate within an Alu retrotransposon element generating RNA transcripts that may have novel functions intracellularly [102].

The impact of a potential cooperation between enhancer (HS2) and B1-559, as proposed in Fig. 5.2b, is very interesting and remains to be further proved. However, the possibility that HS2 distant elements and DNA domains like B1-559 exert a regulatory effect on β^{major} globin gene expression via a "trans-regulated circuit" in hematopoietic cells upon development and under the influence of DNA methylation state is quite challenging. Alternatively, TEs may serve to complex transcriptional switches in eukaryotic systems. One such example applies to the activation of interferon-beta (IFN-β) transcription which has been uncovered to be a highly ordered process beginning with the delivery of NF-kB to the IFN-β enhancer through a process involving stochastic interchromosomal interactions between the IFN-β enhancer and specialized Alu elements [103, 104]. In any case, however, all the above mentioned examples support the hypothesis that TEs in the mammalian genomes might be considered as modulatory elements in transcriptional regulatory circuits (TRCs) ensuring coordinated expression of genes organized in transcriptional units like globin genes and many others.

New Concepts Regarding the Usefulness of "Junk DNA" Toward the Clinical Exploitation of Personal Genome Variations

A few years ago, in the first comprehensive whole-genome analysis of mobile elements-related structural variants of an individual, it has been demonstrated that TEs play an important role in generating interindividual structural variability by estimating the Alu, L1 and SVA retrotransposition rates to be one in 21 births, 212 births, and 916 births, respectively [16]. Now, it is a fortune that very recently a vast amount of functional data from the "ENCODE Project Consortium" (The Encyclopedia of DNA Elements; ENCODE) have been released related to the regulation and function of human and other complex genomes (see at: http://genome.ucsc.edu/ENCODE/ and www.nature.com/encode). It is interesting to note that these data provide new insights into genetic variability patterns seen in individuals and populations especially in terms of 'junk DNA" structure by providing evidence that ~80 % of the human genome serves some function [105–107]. As it has been shown, many previously clinically-validated DNA variants are located outside of the exome and within or very close to intergenic regions and other non-coding functional DNA elements. As a consequence, by also including in genome-wide association studies the variations seen in "junk DNA" could provide new ways on how to more efficiently achieve the clinical translation of genomic information to link specific genetic polymorphisms with disease etiology and progression profiles. Such new genetic information impinge on the regulation of complex mechanisms involved in human genome function which in turn may contribute to molecular pathophysiology mechanisms. To this end, the "junk DNA" functionality in DNA replication, transcription and translation machineries seems to contribute in interindividual variability, more than previously believed, and this in turn may affect the way by which the clinical exploitation of the generated knowledge could be enhanced. By including such information in clinical genomic studies, it is expected to enrich and strengthen our translational medicine capabilities and achieve major benefits in therapeutics for all patients worldwide.

Acknowledgments We would like to thank Prof. Asterios S. Tsiftsoglou (Laboratory of Pharmacology, Department of Pharmaceutical Sciences, Aristotle University of Thessaloniki, Greece) for his thoughtful criticism and continuous support throughout the entire period of our efforts to carry out the experiments, analyze the results and delineate the mechanisms discussed in this chapter.

References

1. Biémont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443(7111): 521–524
2. Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 8:241–259
3. Gu W, Castoe TA, Hedges DJ, Batzer MA, Pollock DD (2008) Identification of repeat structure in large genomes using repeat probability clouds. Anal Biochem 380(1):77–83
4. Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303(5664): 1626–1632
5. Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN (2008) Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res 16(1): 203–215
6. Pace JK, Feschotte C (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res 17(4):422–432
7. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet 2(1):e2
8. Hancks DC, Kazazian HH Jr (2012) Active human retrotransposons: variation and disease. Curr Opin Genet Dev 22(3):191–203
9. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV et al (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A 100(9):5280–5285
10. Ostertag EM, Kazazian HH Jr (2001) Biology of mammalian L1 retrotransposons. Annu Rev Genet 35:501–538
11. Boeke JD (1997) LINEs and Alus – the polyA connection. Nat Genet 16(1):6–7
12. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35(1):41–48
13. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH Jr (2011) Retrotransposition of marked SVA elements by human L1s in cultured cells. Hum Mol Genet 20(17): 3386–3400
14. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M et al (2012) The nonautonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res 40(4):1666–1683

15. Dewannieux M, Heidmann T (2005) Role of poly(A) tail length in Alu retrotransposition. Genomics 86(3):378–381
16. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD et al (2009) Mobile elements create structural variation: analysis of a complete human genome. Genome Res 19(9):1516–1526
17. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA et al (2005) SVA elements: a hominid-specific retroposon family. J Mol Biol 354(4):994–1007
18. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73(6):1444–1451
19. McClintock B (1956) Controlling elements and the gene. Cold Spring Harb Symp Quant Biol 21:197–216
20. Takahashi H, Okazaki S, Fujiwara H (1997) A new family of site specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, Bombyx mori. Nucleic Acids Res 25(8):1578–1584
21. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM (1993) Transposons in place of telomeric repeats at a Drosophila telomere. Cell 75(6):1083–1093
22. Holmes SE, Dombroski BA, Krebs CM, Boehm CD, Kazazian HH Jr (1994) A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nat Genet 7(2):143–148
23. Goodier JL, Ostertag EM, Kazazian HH Jr (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum Mol Genet 9(4):653–657
24. Belancio VP, Hedges DJ, Deininger P (2008) Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res 18(3):343–358
25. Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429(6989):268–274
26. Konkel MK, Batzer MA (2010) A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. Semin Cancer Biol 20(4):211–221
27. van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19(10):530–536
28. Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res 34(5):1512–1521
29. Perepelitsa-Belancio V, Deininger P (2003) RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet 35(4):363–366
30. Mätlik K, Redik K, Speek M (2006) L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol 2006(1):71753
31. Speek M (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21(6):1973–1985
32. Polavarapu N, Mariño-Ramírez L, Landsman D, McDonald JF, Jordan IK (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics 9:226
33. Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R (1998) Apolipoprotein(a) gene enhancer resides within a LINE element. J Biol Chem 273(2):891–897
34. Sharan C, Hamilton NM, Parl AK, Singh PK, Chaudhuri G (1999) Identification and characterization of a transcriptional silencer upstream of the human BRCA2 gene. Biochem Biophys Res Commun 265(2):285–290
35. Gasior SL, Wakeman TP, Xu B, Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol 357(5):1383–1393
36. Pierce AJ, Stark JM, Araujo FD, Moynahan ME, Berwick M, Jasin M (2001) Double-strand breaks and tumorigenesis. Trends Cell Biol 11(11):S52–S59
37. Vilenchik MM, Knudson AG (2003) Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. Proc Natl Acad Sci U S A 100(22):12871–12876
38. Hedges DJ, Deininger PL (2007) Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res 616(1–2):46–59

39. Kang MI, Rhyu MG, Kim YH, Jung YC, Hong SJ, Cho CS et al (2006) The length of CpG islands is associated with the distribution of Alu and L1 retroelements. Genomics 87(5): 580–590
40. Lyon MF (1998) X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet 80(1–4):133–137
41. Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17(6):915–928
42. Phokaew C, Kowudtitham S, Subbalekha K, Shuangshoti S, Mutirangura A (2008) LINE-1 methylation patterns of different loci in normal and cancerous cells. Nucleic Acids Res 36(17):5704–5712
43. Kochanek S, Renz D, Doerfler W (1993) DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J 12(3):1141–1151
44. Rubin CM, Vande Voort CA, Teplitz RL, Schmid CW (1994) Alu repeated DNAs are differentially methylated in primate germ cells. Nucleic Acids Res 22(23):5121–5127
45. Strichman-Almashanu LZ, Lee RS, Onyango PO, Perlman E, Flam F, Frieman MB et al (2002) A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. Genome Res 12(4):543–554
46. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69–87
47. Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13(8):335–340
48. Zamudio N, Bourc'his D (2010) Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Heredity (Edinb) 105(1):92–104
49. Walsh CP, Chaillet JR, Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet 20(2):116–117
50. Maksakova IA, Mager DL, Reiss D (2008) Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. Cell Mol Life Sci 65(21):3329–3347
51. Bourc’his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431(7004):96–99
52. Huang J, Fan T, Yan Q, Zhu H, Fox S, Issaq HJ et al (2004) Lsh, an epigenetic guardian of repetitive elements. Nucleic Acids Res 32(17):5019–5028
53. Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R et al (2009) Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int J Cancer 124(1):81–87
54. Florl AR, Lower R, Schmitz-Drager BJ, Schulz WA (1999) DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas. Br J Cancer 80(9):1312–1321
55. Roman-Gomez J, Jimenez-Velasco A, Agirre X, Cervantes F, Sanchez J, Garate L et al (2005) Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia. Oncogene 24(48):7213–7223
56. Wilson AS, Power BE, Molloy PL (2007) DNA hypomethylation and human diseases. Biochim Biophys Acta 1775(1):138–162
57. Weber B, Kimhi S, Howard G, Eden A, Lyko F (2010) Demethylation of a LINE-1 antisense promoter in the cMet locus impairs met signalling through induction of illegitimate transcription. Oncogene 29(43):5775–5784
58. Garcia-Perez JL, Morell M, Scheys JO, Kulpa DA, Morell S, Carter CC et al (2010) Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature 466(7307):769–773
59. Ronemus M, Martienssen R (2005) RNA interference: methylation mystery. Nature 433(7025):472–473
60. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316(5825): 744–747

61. Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M et al (2008) DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev 22(7):908–917
62. Aravin AA, Bourc'his D (2008) Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev 22(8):970–975
63. Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H et al (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev 20(13):1732–1743
64. Murchison EP, Stein P, Xuan Z, Pan H, Zhang MQ, Schultz RM et al (2007) Critical roles for Dicer in the female germline. Genes Dev 21(6):682–693
65. Golden DE, Gerbasi VR, Sontheimer EJ (2008) An inside job for siRNAs. Mol Cell 31(3):309–312
66. Schumann GG, Gogvadze EV, Osanai-Futahashi M, Kuroki A, Münk C, Fujiwara H et al (2010) Unique functions of repetitive transcriptomes. Int Rev Cell Mol Biol 285:115–188
67. Belancio VP, Deininger PL, Roy-Engel AM (2009) LINE dancing in the human genome: transposable elements and disease. Genome Med 1(10):97
68. Callinan PA, Batzer MA (2006) Retrotransposable elements and human disease. Genome Dyn 1:104–115
69. Solyom S, Kazazian HH Jr (2012) Mobile elements in the human genome: implications for disease. Genome Med 4(2):12
70. Sayah DM, Sokolskaja E, Berthoux L, Luban J (2004) Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature 430(6999):569–573
71. Hamdi HK, Reznik J, Castellon R, Atilano SR, Ong JM, Udar N et al (2002) Alu DNA polymorphism in ACE gene is protective for age-related macular degeneration. Biochem Biophys Res Commun 295(3):668–672
72. Agrawal A, Eastman QM, Schatz DG (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394(6695):744–751
73. Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332(6160):164–166
74. Wimmer K, Callens T, Wernstedt A, Messiaen L (2011) The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. PLoS Genet 7(11):e1002371
75. VandenDriessche T, Ivics Z, Izsvák Z, Chuah MK (2009) Emerging potential of transposons for gene therapy and generation of induced pluripotent stem cells. Blood 114(8):1461–1468
76. Ivics Z, Li MA, Mátés L, Boeke JD, Nagy A, Bradley A et al (2009) Transposon-mediated genome manipulation in vertebrates. Nat Methods 6(6):415–422
77. Mátés L, Chuah MK, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A et al (2009) Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nat Genet 41(6):753–761
78. Ortiz-Urda S, Thyagarajan B, Keene DR, Lin Q, Fang M, Calos MP et al (2002) Stable nonviral genetic correction of inherited human skin disease. Nat Med 8(10):1166–1170
79. Montini E, Held PK, Noll M, Morcinek N, Al-Dhalimy M, Finegold M et al (2002) In vivo correction of murine tyrosinemia type I by DNA-mediated transposition. Mol Ther 6(6):759–769
80. Liu L, Mah C, Fletcher BS (2006) Sustained FVIII expression and phenotypic correction of hemophilia A in neonatal mice using an endothelial-targeted sleeping beauty transposon. Mol Ther 13(5):1006–1015
81. Ohlfest JR, Frandsen JL, Fritz S, Lobitz PD, Perkinson SG, Clark KJ et al (2005) Phenotypic correction and long-term expression of factor VIII in hemophilic mice by immunotolerization and nonviral gene transfer using the Sleeping Beauty transposon system. Blood 105(7): 2691–2698
82. Singh H, Manuri PR, Olivares S, Dara N, Dawson MJ, Huls H et al (2008) Redirecting specificity of T-cell populations for CD19 using the Sleeping Beauty system. Cancer Res 68(8):2961–2971

83. Tsiftsoglou AS, Pappas IS, Vizirianakis IS (2003) Mechanisms involved in the induced differentiation of leukemia cells. Pharmacol Ther 100(3):257–290
84. Tsiftsoglou AS, Pappas IS, Vizirianakis IS (2003) The developmental program of murine erythroleukemia cells. Oncol Res 13(6–10):339–346
85. Vizirianakis IS, Tsiftsoglou AS (1996) Induction of murine erythroleukemia cell differentiation is associated with methylation and differential stability of polyA+ RNA transcripts. Biochim Biophys Acta 1312(1):8–20
86. Vizirianakis IS, Tsiftsoglou AS (1995) N6-methyladenosine inhibits murine erythroleukemia cell maturation by blocking methylation of RNA and memory via conversion to S-(N6-methyl)-adenosylhomocysteine. Biochem Pharmacol 50(11):1807–1814
87. Vizirianakis IS, Tsiftsoglou AS (2005) Blockade of murine erythroleukemia cell differentiation by hypomethylating agents causes accumulation of discrete small poly(A)- RNAs hybridized to 3′-end flanking sequences of beta(major) globin gene. Biochim Biophys Acta 1743(1–2):101–114
88. Vizirianakis IS, Wong W, Tsiftsoglou AS (1992) Analysis of inhibition of commitment of murine erythroleukemia (MEL) cells to terminal maturation by N6-methyladenosine. Biochem Pharmacol 44(5):927–936
89. Vizirianakis IS, Tezias SS, Amanatiadou EP, Tsiftsoglou AS (2012) Possible interaction between B1 retrotransposon-containing sequences and β(major) globin gene transcriptional activation during MEL cell erythroid differentiation. Cell Biol Int 36(1):47–55
90. Tezias SS, Tsiftsoglou AS, Amanatiadou EP, Vizirianakis IS (2012) Cloning and characterization of polyA- RNA transcripts encoded by activated B1-like retrotransposons in mouse erythroleukemia MEL cells exposed to methylation inhibitors. BMB Rep 45(2):126–131
91. Vidal F, Mougneau E, Glaichenhaus N, Vaigot P, Darmon M, Cuzin F (1993) Coordinated posttranscriptional control of gene expression by modular elements including Alu-like repetitive sequences. Proc Natl Acad Sci U S A 90(1):208–212
92. Duncan C, Biro PA, Choudary PV, Elder JT, Wang RRC, Forget BG et al (1979) RNA polymerase III transcriptional units are interspersed among human non-α-globin genes. Proc Natl Acad Sci U S A 76(10):5095–5099
93. Tsiftsoglou AS, Vizirianakis IS, Strouboulis J (2009) Erythropoiesis: model systems, molecular regulators, and developmental programs. IUBMB Life 61(8):800–830
94. Sawado T, Igarashi K, Groudine M (2001) Activation of beta-major globin gene transcription is associated with recruitment of NF-E2 to the beta-globin LCR and gene promoter. Proc Natl Acad Sci U S A 98(18):10226–10231
95. Cantor AB, Orkin SH (2002) Transcriptional regulation of erythropoiesis: an affair involving multiple partners. Oncogene 21(21):3368–3376
96. Zhou Y, Zheng JB, Gu X, Li W, Saunders GF (2000) A novel Pax-6 binding site in rodent B1 repetitive elements: coevolution between developmental regulation and repeated elements? Gene 245(2):319–328
97. Antonaki A, Demetriades C, Polyzos A, Banos A, Vatsellas G, Lavigne MD et al (2011) Genomic analysis reveals a novel nuclear factor-κB (NF-κB)-binding site in Alu-repetitive elements. J Biol Chem 286(44):38768–38782
98. Harris CR, Dewan A, Zupnick A, Normart R, Gabriel A, Prives C et al (2009) p53 responsive elements in human retrotransposons. Oncogene 28(44):3857–3865
99. Weiner AM (2002) SINEs and LINEs: the art of biting the hand that feeds you. Curr Opin Cell Biol 14(3):343–350
100. Chen LL, DeCerbo JN, Carmichael GG (2008) Alu element-mediated gene silencing. EMBO J 27(12):1694–1705
101. Lai CB, Zhang Y, Rogers SL, Mager DL (2009) Creation of the two isoforms of rodent NKG2D was driven by a B1 retrotransposon insertion. Nucleic Acids Res 37(9):3032–3043
102. Romanish MT, Nakamura H, Lai CB, Wang Y, Mager DL (2009) A novel protein isoform of the multicopy human NAIP gene derives from intragenic Alu SINE promoters. PLoS One 4(6):e5761

103. Apostolou E, Thanos D (2008) Virus infection induces NF-kappaB-dependent interchromosomal associations mediating monoallelic IFN-beta gene expression. Cell 134(1):85–96
104. Ford E, Thanos D (2010) The transcriptional code of human IFN-beta gene expression. Biochim Biophys Acta 1799(3–4):328–336
105. Stamatoyannopoulos JA (2012) What does our genome encode? Genome Res 22(9):1602–1611
106. Chanock S (2012) Toward mapping the biology of the genome. Genome Res 22(9): 1612–1615
107. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E (2012) Genomics: ENCODE explained. Nature 489(7414):52–55
108. Shehee WR, Loeb DD, Adey NB, Burton FH, Casavant NC, Cole P et al (1989) Nucleotide sequence of the BALB/c mouse beta-globin complex. J Mol Biol 205(1):41–62

Chapter 6
Copy Number Variation in Human Health, Disease and Evolution

Carolina Sismani, Costas Koufaris, and Konstantinos Voskarides

Introduction

The most widespread type of variation in the human genome is Single Nucleotide Polymorphisms (SNPs). Therefore until recently the basis of studying genome variability and disease pathogenicity was mainly focused on SNPs. However, the advancement of genome-wide technologies that allow comprehensive screening of the entire genome has enabled the identification of another important abundant form of genetic variation in the human genome that of structural variation. These studies have also shown that the extent of structural variation is much greater than was previously anticipated [1].

Structural Variation (SV) is a collective term for a group of genomic alterations that change the structure but not the sequence of the genome. SVs include quantitative changes such as Copy Number Variations (deletions and duplications), positional changes such as translocations and changes in terms of sequence orientation such as inversions. Copy Number Variations (CNVs) represent a large category of structural variation and have been defined as a segment of DNA that is larger than 1 kb and present at a variable copy number in comparison to a reference genome [2].

CNVs can be generally divided into two major categories based on their frequency. The first category includes CNVs which are common in the general population with an overall frequency higher than 1 %. These copy number changes are also

C. Sismani, Ph.D. (✉) • C. Koufaris, Ph.D.
Department of Cytogenetics and Genomics, The Cyprus Institute of Neurology and Genetics,
PO Box 23462, Nicosia, Cyprus 1683
e-mail: csismani@cing.ac.cy; costask@cing.ac.cy

K. Voskarides, Ph.D.
Department of Biological Sciences, Molecular Medicine Research Center,
University of Cyprus, Kallipoleos 75, 1678 Nicosia, Cyprus
e-mail: kvoskar@ucy.ac.cy

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_6

referred as CNPs (Copy Number Polymorphism). Common CNVs (CNPs) are usually small, typically they are less than 10 kb and can have no phenotypic effect on an individual or they might be associated with susceptibility to complex genetic diseases such as psoriasis and Crohn's disease [3, 4].

The second category involves CNVs that are rare in the general population. These CNVs are typically much larger in size than the common CNVs and consequently have a much larger risk of involving dosage-sensitive genes resulting in a phenotypic effect. The pathological conditions caused by genomic rearrangements (deletions/duplications, inversions, insertions and translocations) are collectively defined as genomic disorders [5].

CNV are found in the genomes of all individuals, are wide-spread across the genome, and include both inherited and de novo CNVs [6]. CNVs are usually stable and can potentially be inherited. As the importance of the duplications and deletions that result in these variants is becoming apparent, cataloging them and assessing their frequencies is now an important goal. A continuously updated summary of CNVs can be found in The Database of Genomic Variants (http://projects.tcag.ca/variation/). In addition the Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER; https://decipher.sanger.ac.uk/information) is cataloguing clinically relevant CNVs. Furthermore new genetic disorders/syndromes caused by CNVs are also catalogued in the Online Mendelian Inheritance in Man (OMIM, www.omim.org) database.

Mechanisms of CNV Formation

Three major types of mechanisms have been proposed for the formation of CNVs in the human genome namely, (a) Homologous Recombination Mechanism (HRM) with the major mechanism being Non-Allelic Homologous Recombination (NAHR), (b) Non-homologous recombination mechanism (NHRM) with the major mechanism being Non-homologous End joining (NHEJ) and (c) replication based mechanisms (RBM) with the major mechanism being Fork Stalling and Template Switching/mediated break-induced replication FoSTeS/MMBIR. HRM are mostly associated with recurrent CNVs with recurrent breakpoints and are found in regions of extensive homology (1–5 kb). Non-recurrent rare CNVs occur at regions with very limited homology, microhomology (2–15 bp) or even no homology at all and are mostly formed by NHRM and RBM. These mechanisms occur both in germ cells and somatic cells, where the rearrangements can be associated with genomic disorders and cancer, respectively.

Non-allelic Homologous Recombination (NAHR)

NAHR is the predominant molecular genetic mechanism responsible for recurrent genomic rearrangements, those that share a common size and show clustering of breakpoints. NAHR is mostly mediated by low-copy repeats (LCR)s also known as segmental duplications (SD) which are genomic fragments of high sequence identity (>95 %), usually 10–500 kb in size, which account for ~5 % of the human genome [7].

Meiotic recombination between non-allelic LCRs in direct orientation on the same chromosome results in deletions or reciprocal duplications of the genomic region located between them. The rearrangement breakpoints of NAHR tend to cluster within LCR, termed recombination hotspots.

Non-homologous End Joining (NHEJ)

Some simple non-recurrent rearrangements can occur via Non-homologous end joining (NHEJ) [8]. NHEJ is one of the two major mechanisms employed by the cell to repair double-stranded breaks and has been described in several organisms, from bacteria to mammals. Unlike NAHR, NHEJ does not require substrates with extended homology.

NHEJ involves four steps: detection of double-stranded DNA break, molecular bridging of both broken DNA ends, modification of the ends to make them compatible and ligatable, and the final ligation step.

Microhomology Mediated Replication Error Mechanisms

Recent studies of genomic disorders have shown that complex non-recurrent structural rearrangements, can be explained by replication-based human genomic rearrangement mechanisms such as FoSTeS/MMBIR (fork stalling and template switching/microhomology-mediated break-induced replication) [9, 10]. In these models, the DNA replication fork stalls or collapses, the lagging strand disengages from the original template and anneals to another replication fork in physical proximity, utilizing microhomology at the 3 Ά end, “priming” or reinitiating DNA synthesis [11].

Evolution of CNV

CNVs and Purifying Selection

Generally, CNVs seems to preferentially locate outside of genes and highly conserved elements. It is also quite logical that significantly lower proportion of deletions than duplications overlaps with disease-related genes and more generally with protein coding regions. But are we sure that this large pool of CNVs found in “junk” genetic regions are evolutionarily and phenotypically neutral? No, since these regions are usually understudied in genetic and genomic research. Even in the case of functionally neutral CNVs, this is a quite big pool of amplified DNA that it can potentially gain a role after sequential mutation and/or after change of environmental conditions. We must always have in mind that evolution is a very dynamic procedure and genetic regions that are currently under negative selection can have adaptive benefit/s under a different environment.

Nguyen et al. [12] were the first to support that purifying (negative) selection is the main evolutionary force on CNV regions, despite the fact that the same group gave evidence for the opposite (that positive selection directs CNVs) 2 years before [13]. They finally concluded that positive selection is the exception and not the rule. After separating available by then CNVs in four groups and analyzing each group under an evolutionary concept they concluded in very interesting results. In brief, their data supported a model of reduced purifying selection (Hill-Robertson interference) within copy number variable regions that are enriched in nonessential genes, allowing both the fixation of slightly deleterious substitutions and increased drift of CNV alleles [12]. At the same sense, Schuster-Böckler et al. [14], by analyzing all the available CNV human population database records they found that dosage-sensitive genes are under-represented in CNV regions. They concluded that this is a strong indication for action of negative selection on human CNV regions.

Berglund et al. [15], found evidence for negative selection acting on CNVs in dogs. By analyzing CNV regions in many canine breeds, they noticed that 98 % of them are observed in multiple breeds. CNVs that predicted to disrupt gene function were significantly less common than expected by chance. Their data supported the fact that purifying selection is a major factor acting on structural variation and shaping so the dog genome, concluding from this that many CNVs are "unwarranted" by evolution, even if found in "junk" DNA. These sequence features may have driven genome instability and chromosomal rearrangements throughout canine evolution [15]. In the same line of thinking, recent work by another group support the hypothesis that ohnologs (paralogous genes that have originated by whole-genome duplication) are overrepresented in pathogenic copy number mutations possibly because they include critical dosage-sensitive elements of the genome [16].

Zhou et al. [17] established a number of second-chromosome substitution lines in *Drosophila melanogaster* in order to uncover CNV characteristics when these are in homozygous state. They found that more than 70 % of the dosage-sensitive CNVs are recessive with undetectable effects on transcription in heterozygotes, this being supporting for negative selection effect. Gazave et al. [18] performed comparative genomic hybridizations in four primate species populations in order to reveal CNV frequencies and hotspots. They showed that CNVs fate has been possibly determined by selective pressures in different lineages, this resembling a kind of genomic divergent evolution. Evidence for purifying selection was stronger in gorilla CNVs overlapping genes, while positive selection appeared to have driven the final frequencies of structural variants in the orangutan lineage. In contrast, chimpanzees and bonobos present high levels of common structural polymorphism, which is indicative of relaxed purifying selection.

Gene Duplication and Positive Selection

Very early after CNVs discovery, evidence was found for positive selection acting on these genomic elements, within the modern human populations [13]. Those researchers supported that they found evidence for positive selection by analyzing a set of 627 human CNV regions, based on the fact that those CNVs were enriched in genes, particularly genes implicated in secretion and sensing of the environment.

Poptsova et al. [19] performed pathway enrichment analysis for genes found in CNVs in a collection of individuals with Caucasian, Asian or Yoruban descent, combining high-resolution array and sequencing data. The analysis suggested possible examples of positive selection, in pathways like NF-kB and MAPK signaling, and Alu/L1 retrotransposition factors. The authors believe that their results show that constitutional CNVs may modulate subtle pathway changes through specific pathway enzymes, which may become fixed in some populations [19].

Positive selection may result in pleiotropic effects for the genes in specific CNV regions. An interesting example is the 16p11.2 chromosomal region. Duplications or deletions at the chromosomal locus 16p11.2 have been implicated in microcephaly, autism [20], schizophrenia [21], epilepsy, and other neuropsychiatric disorders, as also in anorexia, underweight and obesity phenotypes [22]. Data show that genetic balance is sensitive at this region. As a result of this, energy balance, brain structure and IQ levels [23] can be easily disturbed by different kind of re-arrangements. Recently, animal model studies showed that the neurological phenotypes at 16p11.2 can be attributed on *KCTD13* gene dosage [24]. It would be interesting if evolutionary studies could be performed for this gene in order to understand better its selection and its role. Human lineage specific traits, like intelligence, may be related with *KCTD13* evolutionary history.

Below we will analyze the three major evolutionary fates of CNVs in genomes and consequently in populations, based on widely accepted models (Fig. 6.1). For a more detailed analysis on these models please read the review of Hahn [25]. Such genetic evolutionary phenomena are difficult to be directly observed since they are happening in the depth of macro-evolutionary time. Despite this, pieces of evidence are always there. Range of genetic diversity is something that can be directly measured and be compared within different species. Sudmant et al. [26] compared the diversity and rates of copy number and single nucleotide variation across the hominid phylogeny. They found a correlation between duplications and single nucleotide diversity, believing that this recapitulates greatly the phylogeny of apes. Duplications are redundant compared with deletions by 2.8-fold. The load of segregating duplications remains significantly higher in bonobos, Western chimpanzees, and Sumatran orangutans-populations that have experienced recent genetic bottlenecks ($P = 0.0014$, 0.02, and 0.0088, respectively). The authors conclude that demographic effects, such as bottlenecks, have contributed to larger and more gene-rich segments being deleted in the chimpanzee lineage and that this effect may have contributed to episodic bursts in CNV during hominid evolution.

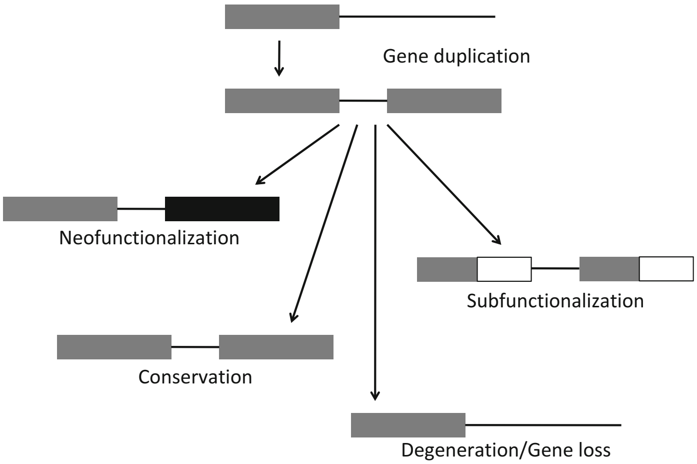


Fig. 6.1 The evolutionary fates of duplicate genes (see text for details)

Neofunctionalization

After the gene duplication, some of the new copy numbers may gain a novel function since they gradually accumulate mutations. This process is called neofunctionalization (Fig. 6.1). Of course, alternatively, if evolutionarily necessary, the new copies can preserve the function of the "maternal" copy for thousands of years (Fig. 6.1). New copies may serve as "back-ups" for this gene function or contribute to increased dosage, if this is required. It is believed that new duplication copies are often free from selective pressure. Thus duplicate genes accumulate mutations faster than a functional single-copy gene. This is why that some scientists believe that this process is the end stage for all subfunctionalized genes.

One characteristic example is the evolution of the antifreeze protein in the Antarctic zoarcid fish. In this case, type III antifreeze protein gene diverged from a paralogous copy of sialic acid synthase (SAS) gene. The ancestral SAS gene was found to have both sialic acid synthase and additional ice-binding properties, this witnessing that neofunctionalization happened. It is obvious here, that after duplication, one of the paralogs began to accumulate mutations that lead to the replacement of the SAS domains of the gene and allowing this way for the antifreeze property to arise. This specialization allows Antarctic zoarcid fish to survive in extreme low temperatures of the Antarctic Sea [27].

Subfunctionalization

Another possible fate for duplicate genes is that both copies are equally free to accumulate degenerative mutations, so long as any defects are complemented by the other copy. This leads to a phenomenon often termed as neutral "subfunctionalization" or DDC (duplication-degeneration-complementation) in which the functionality of the original gene is distributed among the two copies. Neither gene can be lost, as both can perform non-redundant functions, but ultimately neither is able to achieve novel functionality.

A usefulness of the subfunctionalization process is when a gene specializes among different tissues, developmental stages, or environmental conditions, so there is a need of many differently specialized copies. Isozymes are a good example of this phenomenon. These are encoded from paralogs that catalyze the same biochemical reaction, but have evolved a "catalytic freedom", catalyzing a spectrum of substrates in a variety of cell types. Human hemoglobin is an example of a subfunctionalization process, despite this is not a typical enzyme. The ancestral gene copy is undoubtedly a version of the beta globin gene. Through the subfunctionalization, new versions of the gene derived, like the alpha globin genes. Today, functional hemoglobin molecules are dimers of a number of available chains. For an analytical review on evolution of mammalian globin genes please read Storz et al. [28].

One other good example of segregation avoidance occurs in the acetylcholinesterase (AChE1) locus of the common mosquito, Culex pipiens. An allele of this gene has evolved to confer resistance to organophosphate pesticides. It has been found in heterozygote form as a separate duplicated locus in multiple mosquito populations (Hahn [25] and references therein).

Loss

Sometimes, genomic copy numbers can lead to deleterious increased gene expression such as Rett-like syndrome and Pelizaeus–Merzbacher disease [29]. Such pathogenic mutations are likely to be lost from the population and it is very unlikely to be established in the population and gain new functions. However, it is widely known that most duplications are in fact not damaging or beneficial. These neutral sequences can be increased or lost in the population randomly by genetic drift, this being an evolutionary process in molecular level that Kimura proposed for first time in 1962 (neutral evolution).

The Examples of the *DUF1220* and the Amylase Genes

DUF1220 domains' duplications are of great evolutionary interest in human speciation. DUF1220 protein domains have been duplicated many times in the human lineage probably exhibiting the most extreme human lineage–specific copy number increase of any protein coding region in the human genome [30, 31]. The majority of DUF1220 sequences are located at 1q21.1. Copy numbers at the chromosomal region 1q21 have been recently associated with neurological disorders and human cognitive functions as well. Neurological disorders that have been associated with this region are microcephaly, autism and schizophrenia. Teams of Dumas and Sikela showed that copy numbers of DUF1220 are highly expanded in humans, reduced in African great apes, further reduced in orangutan and Old World monkeys, represent only a single-copy in non-primate mammals, and are absent in non-mammalian species. Examination of expression in brain showed that neuron-specific DUF1220 signals are present in the cortical layers of the hippocampus and they are also abundant in neurons within the neocortex. Davis et al. [32] showed a positive correlation between the DUF1220 copy numbers and the human cognitive abilities in two independent populations of European descent, this being the first replicated association with these traits. The actual cognitive traits measured were the total IQ and mathematical aptitude scores. It is obvious through these and other studies, that expansion of DUF1220 copy numbers follow the expansion of the human brain size and the development of the higher cognitive functions in apes.

The α-amylase genes are located in a cluster on the chromosome that includes salivary amylase genes (*AMY1*), two pancreatic α-amylase genes (*AMY2A* and *AMY2B*) and a related pseudogene. The *AMY1* genes show extensive copy number variation which is directly proportional to the salivary α-amylase content in saliva. Perry et al. [33] published an exceptional paper examining the increase in copies of the salivary amylase gene in humans throughout evolutionary history. Perry and colleagues found that *AMY1* copy numbers started to increase 120,000 years ago, something that was catalytic for adaptation of humans to starch consumption. A most recent study found very strong evidence that *AMY1* copy number are associated with increased Body Mass Index (BMI) [change in BMI per estimated copy $=-0.15$ (0.02) kg/m^2; P $=6.93\times10^{-10}$] and obesity risk [odds ratio (OR) per estimated copy $=1.19$, 95 %, CI $=1.13-1.26$; P $=1.46\times10^{-10}$]. This is one of the many evolutionary examples where we observe a change of gene function after a sudden environmental change. Our genome has been evolved under very different environmental conditions than today and evolutionary time was not adequate to adopt in food plethora of present times. This is part of the answer why today we have an extreme increasing of diseases like obesity and diabetes.

Co-evolution of CNVs with MicroRNAs

Evolution of genomic elements must be always considered in relation with other genomic elements, especially the ones having a special regulatory role. This subchapter summarizes recent findings about co-evolution of CNVs with microRNAs, that it is discussed in more detail in Chap. 1.

Existing data show that tandem duplications can result in paralogous microRNA sequences that are located on the same transcript and are organized as tandem paralog clusters [34, 35]. In a recent study, evidence was found that repetitive elements contribute to creation of new microRNAs in mammals and that large segmental duplication events accelerate the expansion of microRNA families, including those derived from repetitive sequences [36]. The latter ones are considered as the younger microRNA genes, being also the less conserved. Similar evidence was found in plants where microRNAs that are found in repetitive elements tend to have longer hairpin precursor, lower G-C content in hairpin precursor sequences and lower minimum free energy [37].

An interesting model for the deriving of new microRNA genes has been proposed by Allen et al. [38], called the inverted duplication model. Under this hypothesis, new microRNA genes are generated from inverted duplication events happened on one of their target genes by forming two adjacent gene segments in either convergent or divergent orientation. Recent observations showed that many microRNA genes are found in transposable elements and pseudogenes. This was considered as an indication that these microRNAs have been derived through inverted duplications. Zhang et al. [35] confirmed further the inverted duplication model in plants, this happening via TEs or pseudogenes, showing also that inverted duplications give rise to microRNAs much more frequently that segmental duplications.

A crucial matter in microRNAs evolution is what factors determine the type and the number of microRNAs' targets in 3′ UTRs. Ha et al. [39] proved that small RNAs produced during interspecific mating or polyploidization serve as a buffer against the genomic shock in interspecific hybrids and allopolyploids. The authors came to this conclusion after studying allotetraploids coming from *A. thaliana* and *A. arenosa*, identifying adoptive alterations of the microRNAs and siRNAs levels in comparison with the parental species. Abrouk et al. [40] found evidence that the above mechanism may be a standard procedure in plants after euploidy, especially in euploidy events that are involved in evolutionary speciation.

But since whole genome duplications are very rare to animal species, research has been performed for duplications of smaller scale in such species. Li et al. [41] and D'Antonio and Ciccrelli [42] found that microRNA targets are significantly enriched for paralogs genes. Characteristically, Li et al. [41] mention that their results suggest that "microRNA-mediated regulation plays an important role in the regulatory circuits involving duplicated genes including adjusting imbalanced dosage effects of gene duplicates, and possibly creating a mechanism for genetic buffering". A more complicate analysis by Fernandez and Chen [43] revealed that human paralogs of poorly packed proteins (categorized so according special structural criteria) are more likely to be targeted by microRNAs, thus underscoring a means to buffer dosage imbalance effects arising from gene duplication.

Recently, our team proceeded to analysis of available data of public genetic databases and we found evidence that miRNAs and Copy Number Variations must have co-evolved and interacted in a way to maintain the balance of the dosage sensitive genes. Our findings raised the possibility that miRNAs may have been created under evolutionary pressure, as a mechanism for increasing the tolerance to genome plasticity [44, 45]. Our results were further confirmed by Woodwark and Bateman [46]. There is significant indication that a number of genes located within CNVs have altered expression level [9, 47–49]. This suggests further that microRNAs may have acted as equilibrators of gene expression during evolution in an attempt to regulate aberrant gene expression and to increase the tolerance to genome plasticity.

Methodologies for Detecting CNVs

The discovery of CNVs has been accelerated in the last years due to the advances in array-based methods and more recently Next Generation Sequencing (NGS) that allow comprehensive screening of the entire genome. In addition several other methodologies have been implemented that target defined loci such as Fluorescence In situ Hybridization (FISH), Quantitative Polymerase Chain Reaction (Q-PCR) or other probe based methods like Multiplex Ligation Probe Amplification (MLPA) and Multiplex Amplifiable Probe Hybridization (MAPH).

Whole Genome Approaches

Array-Based Methods

Two different array-based approaches are currently used for the detection and analysis of CNVs. The first approach is based on array-Comparative Genomic Hybridization (array-CGH) using differentially labelled reference and test samples which are co-hybridized on an array slide containing mapped DNA sequences which are spaced across the whole genome. This method was initially described in 1997 but has tremendously evolved over the years [50]. Today the most widely used DNA sequences for mapping DNA are BAC clones or short oligonucleotides [51, 52].

In array-CGH, the relative fluorescent intensities of two DNA samples (tagged with two different fluorophores respectively) hybridized to glass attached probes is measured. If there is more fluorescent signal from the sample under investigation relative to the control, then there is a gain of that specific genomic region, and if there is less signal relative to the control, then there is a loss of that DNA region. Gains and losses of clones are detected as a ratio that is plotted against the annotated genomic position (Fig. 6.2).

The resolution of array-CGH is based on the density, size and genomic distance of the probes. Currently commercially available arrays contain up to one million oligonucleotide probes (Agilent technologies, www.agilent.com).

The second approach utilizes the existing SNP (Single Nucleotide Polymorphism) genotyping platforms [53]. Although both approaches are quite similar there is a fundamental difference between the two approaches, as SNP arrays were initially designed for genotype and later validated for CNV detection. The difference between the two platforms are in the type of hybridization, comparative hybridization, comparing it to a reference sample versus single one colour hybridization following a subsequent comparison with a set of reference values from controls. In addition, genotyping arrays provide information on SNP genotypes. The resolution of the SNP array is based on the number of SNPs on the array. Currently commercially available SNP arrays contain well over one million SNPs.

In the last few years, SNP-CGH "hybrid arrays" combining properties from both platforms have been developed (CGH-SNP arrays) and include probes in regions with known copy number variation that do not contain SNPs (Affymetrix, www.affymetrix.com and Illumina, www.illumina.com).

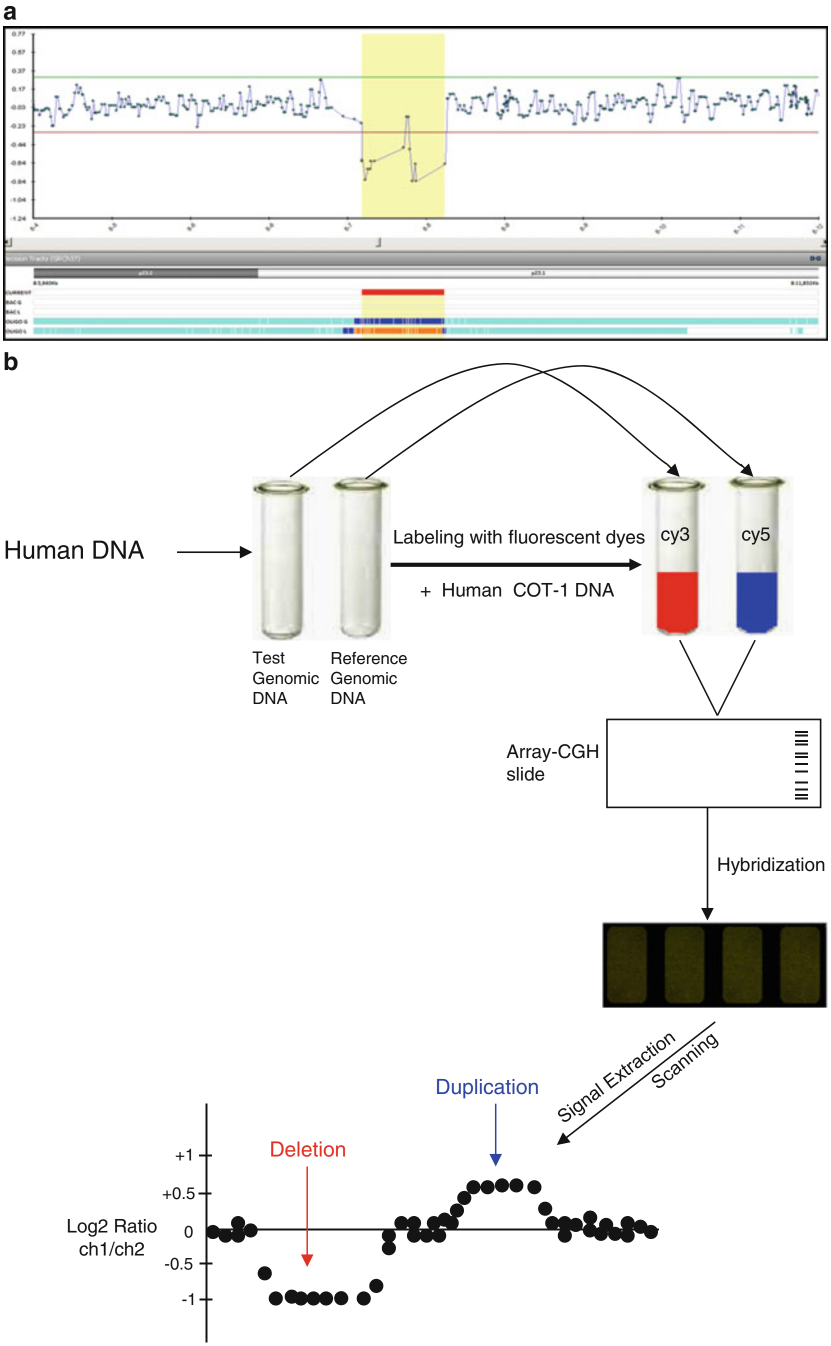


Fig. 6.2 (**a**) Illustration of a deletion (*highlighted*) of approximately 0.8 Mb in size on the short arm of chromosome 8 (q-arm) at chromosomal band 8p23.1 [(location: 7242978–8079890) using build GRCh37 (hg19)]. The current deletion lies within a highly polymorphic region as it is indicated shown by the *dark blue* for duplication and *orange* for deletion. The colors represent the recurrence of the same aberration regarding to our local database. (**b**) Array-CGH methodology for the identification of CNV

Next Generation Sequencing (NGS)

Traditionally, array-CGH is the golden standard for genome-wide method of choice detection of CNVs. However, as resolution of array-CGH is relatively low, higher than 10 kb, even for the ultrahigh resolution platforms of one million probes, smaller CNVs are still difficult to detect. In the last years, NGS, a high throughput sequencing technology has revolutionized the field of genetics [54]. Whole genome NGS has now emerged as a very promising tool for genome-wide detection of CNVs at much higher resolution than array-CGH high resolution (<10 kb) [55]. In addition, NGS has the benefit to detect any type of variation even down to the base pair level [56]. In NGS sequence assemblies from test and reference samples are compared computationally and differences in sequences, copy number and orientation are annotated. However as statistical approaches of NGS are still very limited in relation to CNV detection, further studies are needed prior to its routine implementation for the detection of CNVs.

Targeted/Locus Specific Approach

Fluorescence In Situ Hybridization (FISH)

FISH is a very well established method with a wide range of applications in cytogenetics. FISH is a targeted locus specific approach to detect CNVs. FISH employs fluorescently labeled probes that are hybridized on metaphase chromosomes from the test samples and hybridization patterns (absence/presence, duplication of signal) are compared to reference samples. However the low resolution (>100 kb) as well as the targeted nature of the method does not allow FISH to be implemented on large scale genome wide screenings.

Q-PCR

Real Time quantitative PCR (RT-PCR or Q-PCR) is another targeted strategy for the detection of CNVs. The basis of Q-PCR is that the rate of amplification of a region is proportionate to the number of template copies [57, 58]. Amplification (fluorescence signal representing the target) is then measured in real time during exponential phase and compared to a control region of known copy number, allowing the detection of copy number of the target region.

Probe Based Multiplex Assays

Multiplex Ligation-dependent Probe Amplification (MLPA) and Multiplex Amplifiable Probe Hybridization (MAPH) are able to detect abnormal copy numbers based on quantification of probes specifically designed for the regions of interest [59, 60]. Both methods are based on multiplex PCR methodology using fluorescently labeled universal primers and rely on probe amplification of different size probes rather than amplification of the genomic test DNA. Both methods can detect abnormal copy numbers for up to 50 regions simultaneously and hybridization ratios are compared with control regions. The main difference between the two methods is that MLPA is performed in solution and the probes are designed in two parts and are dependent on the presence of the targeted region for ligation and creation of a contiguous probe to be amplified. In MAPH the denatured test DNA is bound on a nylon membrane and then hybridized with a probe set which are subsequently amplified.

Again, the targeted nature of both methods do not allow implementation on large scale genome wide screening.

Effect of CNV on Human Variation and Disease

Mechanisms by Which CNV Affect Cell or Organism Phenotype

The most widely accepted mechanism by which CNV, whether residing in coding or "junk" DNA, can affect the phenotype of cells or organisms is by altering the expression levels of dosage-sensitive genes. For the majority of human genes modest increases or decreases in expression levels caused by CNV deletions or duplications will have minimal or no observable phenotypic effects. However, certain classes of genes are known to be particularly sensitive to dosage levels changes. The best established categories of dosage sensitive genes are those coding for structural proteins or members of protein complexes [61]. Theoretical and experimental data supports that modest changes in the dosage levels of such genes can disrupt protein–protein interactions and assembly of molecular complexes, thus impairing their biological functions. Similarly, modest changes in copy number of rate-limiting enzymes in metabolic pathways can also affect the output and function of metabolic and signalling pathways.

A straightforward mechanism by which CNV can affect gene dosage is by altering the number of gene copies in the genome. Studies have demonstrated that CNV deletions or duplications of protein-coding genes are in general positively correlated to changes in gene expression [48, 62]. Additionally, CNV located in non-coding regions of the DNA are also known to be able to affect expression of genes. One way this can be achieved is through the disruption of *cis* and *trans*-regulatory sequences (such as enhancers, insulators, and promoters), thus affecting the expression of genes that are under the regulation of these elements. It has also been determined that large CNV can affect the expression of neighbouring genes outside the affected region, as well as in genomic regions further away [63, 64]. Changes in the copy number of non-coding regulatory RNA genes (such as microRNA and long non-coding RNA) can also disrupt the physiological regulation of their target genes. For example CNV amplifications of miRNA will result in lower expression of their target genes, while conversely CNV deletions of the same miRNA gene will result in their increased expression.

Dramatic over-expression of individual genes can also result in prominent phenotypic effects which are not observable after modest changes in gene expression. For example the dramatic amplification of oncogenes such as tyrosine kinases, growth factor receptors, and cell growth genes—often to hundreds of copies per genome—can be a major driver of the process of carcinogenesis. Highly over-expressed proteins can also engage in promiscuous off-target molecular interactions, affecting phenotypes that are irrelevant to the normal functioning of the protein [65].

Complete loss-of-function can occur when CNV deletions unmask a recessive mutation or epigenetic aberration affecting the second allele. The complete loss-of-function of genes with non-redundant biological functions due to the combination of a CNV deletion with genetic or epigenetic lesion on the second allele can adversely affect the fitness of the organism. The aberration affecting the second allele can be another CNV, a point mutation or dysfunctional epigenetic regulation. For example, the coexistence of CNV deletions with point mutations affecting NRXN1 or CNTNAP2 genes has been found in patients with Pitt-Hopkins-like syndrome [66]. A single CNV deletion can also result in loss-of-function of genes if located in the X chromosome in males or when one allele is silenced by parental imprinting mechanisms.

Researchers have also suggested mechanisms by which large-scale CNV can influence organism fitness irrespective of the effects on the individual genes included within the affected genomic regions. A first proposed mechanism is that the simultaneous over-expression of a large number of proteins can act to overwhelm protein control mechanisms and proteasomal degradation, potentially impacting essential cellular functions [67]. A second proposed mechanism is that the large scale amplification of genomic regions containing repetitive elements acting as a drain on the cellular methyl pool, thus promoting DNA hypomethylation and associated genomic instability [68].

Contribution of CNV to Non-pathological Individual Variation

On average CNV affect around four million bases per genome, which is similar to the number of bases that differ between individuals due to single nucleotide polymorphisms (SNP) [69]. Consequently, CNV could be important contributors to the "normal", non-pathogenic phenotypic variability between humans. At present much less attention has been placed on CNV underlying non-pathogenic human traits, although some interesting links have been reported. One interesting case discussed previously relates to the salivary amylase gene ("The Examples of the *DUF1220* and the Amylase Genes" section above). In regards to polygenic complex human traits, studies have indicated that CNV are involved in the observed variability in intelligence [70, 71], height [72, 73], and even musical aptitude [74], although these remain highly debated.

Interpretation of the Pathogenic Significance of CNV

The demarcation of benign from pathogenic CNV is a challenging, but also an essential step for the accurate interpretation of clinical and research findings. Criteria used for the interpretation of CNV include its inheritance pattern, frequency in the population, genomic size of the CNV, and gene content. Classical evolutionary theory predicts that purifying selection acts to remove inherited CNV that are detrimental to the organism from the gene pool (see section "Evolution of CNV" for details). Indeed, CNV disrupting protein-coding genes are depleted amongst high frequency deletions [69], indicating the action of purifying selection. It can therefore be predicted based on evolutionary theory that an inverse association should exist between the penetrance of a given CNV and its frequency in the population. It is expected that common CNV are benign or have very low to moderate penetrance and contribute only modestly to disease risk, while pathogenic CNV of high penetrance are present at lower frequencies in the population. Determining whether a CNV found in a patient was of de novo origin or is inherited from a healthy parent and investigating the presence of the variant in databases of patient cohorts (e.g. DECIPHER decipher.sanger.ac.uk) or healthy controls (e.g. DGV dgv.tcag.ca) are important steps in CNV evaluation, especially for severe phenotypes.

CNV in Intellectual Disability/Developmental Delay and Microdeletion/Microduplication Syndromes

Among the earliest and best established pathogenic phenotypes linked to CNV are intellectual disability and developmental delay. It had been established for a long time by traditional karyotyping that chromosomal abnormalities are associated with intellectual disability and development delay. Over the past decade large case-control studies indicate that 15–20 % of ID and developmental delay cases are caused by CNV that were too small to be detected by traditional karyotyping [75, 76]. Screening for pathogenic CNV is now routinely used for clinical diagnosis of ID and developmental delay.

Microdeletions and microduplications have been recognized as a causative genetic factor of syndromic intellectual disability for many decades now. Typical examples include Prader–Willi (OMIM# 176270) and Angelman (OMIM# 105830) syndromes (15q11-q13 deletion), Williams–Beuren syndrome (7q11.23 deletion, OMIM# 194050), Smith–Magenis syndrome (deletion of 17p11.2, OMIM# 182290) and others. Prior to the implementation of genome wide array-CGH and the establishment of publically available databases, the identification of these syndromes was based on genotype-phenotype approach where a series of patients with similar recognizable clinical features were investigated and the genetic cause of the syndrome was subsequently discovered. However many syndromes have a wide range of clinical features, variability in expression and penetrance which hampered this genotype-phenotype approach. Implementation of high-resolution genome wide array-based approaches, in the last decade, has led to the identification of continuously growing number of recurrent microdeletion and microduplication syndromes caused by submicroscopic CNVs such as the 1p36 microdeletion syndrome, the 17q21.31 microdeletion syndrome and many others [77, 78]. Whole genome array-CGH investigation allows the identification of a similar genomic aberration in patients prior to common clinical presentation delineation. Today more than hundred new microdeletion/microduplication syndromes have been identified [79] using both approaches.

As most recurrent genomic rearrangements are mediated by sequences such as segmental duplications [80] and low copy repeats and are caused by NAHR it was expected that the implementation of array-CGH would lead to the identification of the reciprocal duplication of most of the previously identified recurrent microdeletion syndromes. Indeed, much reciprocal duplication was identified such as the 7q11.23 microduplication syndrome and the 22q11.2 microduplication syndrome and many others [81–83]. However the reciprocal duplication were identified for only a fraction of the microdeletion deletion syndromes mostly due to the fact that microduplications generally result in milder phenotypes or sometimes in no phenotype and consequently can escape detection.

Furthermore, the collection of clinical and genetic information in databases such as DECIPHER and other free databases has been crucial for discriminating between patients with rare aberrations and those with new microdeletion/duplication syndromes.

CNV in Psychiatric Diseases

It is now established that pathogenic CNV are important genetic contributors to the incidence of major psychiatric diseases such as autism, schizophrenia, and bipolar disorder. Particularly strong evidence supports causal associations between rare and de novo CNV and psychiatric diseases, with a number of studies reporting an enrichment of these CNV types in psychiatric patients compared to controls [84]. Beyond their enrichment in psychiatric patients, evidence that CNV are implicated in these diseases has also been gathered from animal models. One example is the display of autism-related behaviour in mice knockout for Cntnap2, a neuronal transmembrane protein implicated in autism [85]. In animal models CNV have also been demonstrated to cause similar anatomical abnormalities as observed in humans, for example 16p11.2 affecting brain growth (for details see “Gene Duplication and Positive Selection” section above) [20].

Ultimately, the CNV disruption of susceptibility genes will in some manner interfere with the physiological development of the brain. The strong links between CNV and psychiatric diseases possibly reflect the inherent sensitivity of neurodevelopmental pathways and neuronal structures involved in learning and memory to abnormal gene dosage. In accordance with this a significant number of established susceptibility genes affected by CNV are synaptic scaffolding proteins or regulators of the levels of synaptic proteins [86]. A significant confounding factor for interpreting the role of CNV in psychiatric diseases is their variable expressivity and incomplete penetrance [84]. Almost all CNV associated with psychiatric disorders are present at lower frequencies in neurotypical controls, while individual CNV are often associated with a number of psychiatric diseases. This phenotypic variability can be probably attributed to the interaction of the CNV with environmental factors and with other genetic factors in order to generate the pathological phenotype. A second issue is that recurrent CNV in psychiatric diseases are often very large, making it difficult to locate the pathogenic genes. An important challenge for researchers is going to be to elucidate how CNV cause the neurobiological abnormalities that eventually lead to the development of psychiatric diseases.

CNV in Cancer

CNV, of both inherited and de novo origin, have been associated with an increased disposition to cancer, prognosis of the disease, and response to therapies. Rare inherited CNV have been detected in nearly half the known cancer predisposition genes, potentially contributing to the incidence of human cancers [87, 88]. However, highly penetrant inherited CNV are believed to account for only a small fraction of familial cancers, which themselves constitute a minority of clinical cancer cases. Consequently, it is considered probable that CNV of low or moderate penetrance are the main contributors to hereditary cancer predisposition associated with CNV [88]. The identification of CNV of low or moderate penetrance associated with increased predisposition to cancer is a challenging task, requiring very large cohorts. Nevertheless, in recent years studies have reported a number of common CNV to be associated with increased disposition to various cancers, for example CNV in 20p13 [89], 2p24.3 [90], and CNV-67048 [91]. An interesting case of an inherited CNV deletion within a gene desert that was shown to be associated with a 1.3 odds ratio of pancreatic cancer is CNVR2966.1 at 6q13 [92]. The proposed mechanism by which this inherited CNV affects cancer risk is its ability to directly interact with the upstream sequence of *CDKN2B* and affect the expression of this known cancer gene. It is expected that efforts to identify inherited CNV associated with increased cancer risk will be an area of intensive research in the coming years.

The accumulation of somatic mutations in the form of aneuploidies, point mutations, and CNV is an essential aspect of carcinogenesis and neoplastic progression. Studies have demonstrated that somatic CNV are especially prominent across cancer genomes [93]. The high rate of de novo CNV formation in tumor tissue is in accordance with the genomic instability and high mutation rate that characterises cancer cells. Recently the largest study examining the patterns of somatic CNV across various cancer types has been published [94]. In this study the authors examined the copy number profiles of 4943 primary cancer specimens across 11 cancer types. The authors report patterns of somatic CNV that were common in the various cancers examined. One observation was the tendency for CNV to be longer and more frequent near the telomeres compared to CNV located internally, indicating that these are formed by different mechanisms. By analysing all cancer lineages this study identified 70 recurrently amplified and 70 recurrently deleted genomic regions, however only a minority of those regions contained known cancer genes. Another observation was that large duplications of genome (such as whole genome duplication) are associated with subsequent increases in somatic CNV within those regions. It is established that CNV disruptions of cancer genes contribute to tumorigenesis. However, a large number of somatic CNV detected in tumors are expected to be "passenger" mutations that arise during cancer evolution but do not contribute to the malignancy. The demarcation of driver from passenger CNV in tumors is a highly challenging task. The strongest case for somatic CNV contributing to cancer development can be made for those that recur at appreciable frequencies, indicating that they contribute to cancer development. However, recurrent CNV could also be explained in certain cases due to their location in fragile sites or due to absence of negative selection against them in evolving cancer genomes.

The classic Knudson's "two-hit" model considers that cancer progression requires the disruption of both alleles of tumor-suppressor genes. According to this model, a CNV deletion, either of germline or somatic origin, resulting in the loss-of-function of one copy of a tumor-suppressor gene is only adverse when the second functioning allele is also disrupted. The disruption of the second allele can be another CNV deletion, a point mutation or epigenetic dysregulation. Germline CNV deletions of tumor suppressor genes can increase the predisposition against cancer, as only a single mutational event that affects the second allele is required to complete the loss-of-function of the protein. Somatic CNV deletions of tumor suppressor genes such as Tp53 and retinoblastoma have been long established as important mutational events in the multi-step process of carcinogenesis, again requiring the disruption of second allele to allow cancer progression. Recently an alternative model to the "two-hit" model has been proposed to explain the recurrent focal regions of hemizygosity observed in cancer which do not contain known tumor-suppressor genes [95]. According to this model the synergistic haploinsufficiency of genes involved in repressing cell proliferation result in a proliferative advantage to cancer cells.

Unlike tumor-suppressor genes, oncogenes drive carcinogenesis through their increased activity. Somatic amplification of gene copy number of oncogenes has long being accepted to be an important carcinogenic mechanism. Classic examples include the amplification of Myc in Burkett's lymphoma and N-Myc in neuroblastoma. More recent studies have highlighted that CNV affecting regulatory elements can also drive oncogenic expression. A recent example of this is the finding of recurrent structural variants in medulloblastomas placing the *GLI1* family genes close to active enhancers, increasing their expression, which contributes to cancer progression [96]. Although less established than CNV affecting tumor suppressor genes, inherited CNV affecting oncogenes are also implicated in cancer. For example individuals with higher copy number of the g.CNV-30450 CNV located in the promoter region of the MAPKAPK2 oncogene resulting in greater expression of the protein, had greater incidence and worse prognosis of lung cancer [97].

Genomic instability is often observed in cancer. The increased rate of formation of new mutations is believed to be a driver of tumorigenesis by causing new carcinogenic mutations. Disruption of genes involved in DNA repair and maintenance can also lead to a greater mutation rate and susceptibility to cancer development. Subset of rare CNV associated with cancer, affect genes with genomic stability and DNA repair such as BRCA1, MSH2, and HRPT2 [88]. At least one example of a common CNV associated with greater mutation rates has been recently reported, affecting the cytidine deaminase APOBEC3B [98], with perhaps many more remaining to be identified.

CNV and Xenobiotics Metabolism

Humans possess highly specialised enzymatic machinery involved in the biochemical modification and excretion of substances that are of external origin. Inherited CNV affecting the number of copies of xenobiotics enzymes were amongst the first to be recognised and have been linked to individual susceptibility to environmental agents and to pharmaceuticals. These inherited CNV are also of particular interest due to their high degree of variability between individuals and ethnic populations.

Evidence supports that CNV affecting metabolic enzymes involved in the detoxification and removal of environmental carcinogens are associated with cancer incidence. The most extensively studied are CNV deletion polymorphisms affecting the *GSTM1* gene, which codes for an phase II metabolism enzyme involved in the glutathione-mediated reduction of electrophilic chemicals, which has been linked by molecular epidemiology studies to cancer susceptibility [99]. Another interesting case involved CNV affecting *CYP2A6*, an enzyme with a crucial function in nicotine metabolism and clearance. The number of copies of *CYP2A6* varies between ethnicities and has been associated with smoking behaviour and tobacco-related diseases [100].

CNV also affect the efficacy and side-effects associated with pharmaceutical treatments. *SULT1A1* is an enzyme catalyzing the sulphate conjugation of a wide variety of drugs. Deletions and duplications of this gene (ranging from one to five copies) are correlated with the hepatic activity of this enzyme, supporting the functional effects of CNV affecting this gene. Again, considerable inherited variability in copy number exists between populations, with 26 % of Caucasians vs. 63 % of African-Americans having three or more copies of *SULT1A1* [101]. Another important potential link is an association between the *CYP2D6* genotype and response to tamoxifen in postmenopausal women, although this is currently a matter of debate [102].

Other Diseases Reported to Be Affected by CNV

Beyond the more established contribution of CNV to intellectual disability, psychiatric diseases, and cancer, studies have indicated the contribution of CNV to a wide range of other human diseases. Examples include CNV linked to risk of osteoporosis [103], early-onset obesity [104], atherosclerosis [105], and Alzheimer's disease [106]. These will also be areas of intensive research in the future.

References

1. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64
2. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97
3. Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, van de Kerkhof PC, Traupe H, de Jongh G, den Heijer M, Reis A, Armour JA, Schalkwijk J (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40:23–25
4. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79:439–448
5. Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14:417–422
6. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H et al (2006) Global variation in copy number in the human genome. Nature 444:444–454
7. Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. Trends Genet 18:74–82
8. Lieber MR, Ma Y, Pannicke U, Schwarz K (2003) Mechanism and regulation of human non-homologous DNA end-joining. Nat Rev Mol Cell Biol 4:712–720
9. Slack A, Thornton PC, Magner DB, Rosenberg SM, Hastings PJ (2006) On the mechanism of gene amplification induced under stress in Escherichia coli. PLoS Genet 2:e48
10. Lee JA, Carvalho CM, Lupski JR (2007) A DNA replication mechanism for generating non-recurrent rearrangements associated with genomic disorders. Cell 131:1235–1247
11. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41:849–853
12. Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP (2008) Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Res 18:1711–1723
13. Nguyen DQ, Webber C, Ponting CP (2006) Bias of selection on human copy-number variants. PLoS Genet 2:e20
14. Schuster-Bockler B, Conrad D, Bateman A (2010) Dosage sensitivity shapes the evolution of copy-number varied regions. PLoS One 5:e9474

15. Berglund J, Nevalainen EM, Molin AM, Perloski M, LUPA Consortium, André C, Zody MC, Sharpe T, Hitte C, Lindblad-Toh K, Lohi H, Webster MT (2012) Novel origins of copy number variation in the dog genome. Genome Biol 13:R73
16. McLysaght A, Makino T, Grayton HM, Tropeano M, Mitchell KJ, Vassos E, Collier DA (2014) Ohnologs are overrepresented in pathogenic copy number mutations. Proc Natl Acad Sci U S A 111:361–366
17. Zhou J, Lemos B, Dopman EB, Hartl DL (2011) Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster. Genome Biol Evol 3:1014–1024
18. Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, Baker C, Marquès-Bonet T, Eichler EE, Navarro A (2011) Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Res 21:1626–1639
19. Poptsova M, Banerjee S, Gokcumen O, Rubin MA, Demichelis F (2013) Impact of constitutional copy number variants on biological pathway evolution. BMC Evol Biol 13:19
20. Horev G, Ellegood J, Lerch JP, Son YE, Muthuswamy L, Vogel H, Krieger AM, Buja A, Henkelman RM, Wigler M, Mills AA (2011) Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. Proc Natl Acad Sci U S A 108:17076–17081
21. Guha S, Rees E, Darvasi A, Ivanov D, Ikeda M, Bergen SE, Magnusson PK, Cormican P, Morris D, Gill M, Cichon S, Rosenfeld JA, Lee A, Gregersen PK, Kane JM, Malhotra AK, Rietschel M, Nöthen MM, Degenhardt F, Priebe L, Breuer R, Strohmaier J, Ruderfer DM, Moran JL, Chambert KD, Sanders AR, Shi J, Kendler K, Riley B, O'Neill T, Walsh D, Malhotra D, Corvin A, Purcell S, Sklar P, Iwata N, Hultman CM, Sullivan PF, Sebat J, McCarthy S, Gejman PV, Levinson DF, Owen MJ, O'Donovan MC, Lencz T, Kirov G, Molecular Genetics of Schizophrenia Consortium, Wellcome Trust Case Control Consortium 2 (2013) Implication of a rare deletion at distal 16p11.2 in schizophrenia. JAMA Psychiatry 70:253–260
22. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, Martinet D, Shen Y, Valsesia A, Beckmann ND, Thorleifsson G, Belfiore M, Bouquillon S, Campion D, de Leeuw N, de Vries BB, Esko T, Fernandez BA, Fernández-Aranda F, Fernández-Real JM, Gratacòs M, Guilmatre A, Hoyer J (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478:97–102
23. Zufferey F, Sherr EH, Beckmann ND, Hanson E, Maillard AM, Hippolyte L, Macé A, Ferrari C, Kutalik Z, Andrieux J, Aylward E, Barker M, Bernier R, Bouquillon S, Conus P, Delobel B, Faucett WA, Goin-Kochel RP, Grant E, Harewood L, Hunter JV, Lebon S, Ledbetter DH, Martin CL, Männik K, Martinet D, Mukherjee P, Ramocki MB, Spence SJ, Steinman KJ, Tjernagel J, Spiro JE, Reymond A, Beckmann JS, Chung WK, Jacquemont S, Simons VIP, Consortium, 16p11.2 European Consortium (2012) A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. J Med Genet 49:660–668
24. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, Reymond A, Sun M, Sawa A, Gusella JF, Kamiya A, Beckmann JS, Katsanis N (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature 485:363–367
25. Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 100:605–617
26. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J, Great Ape Genome Project, Marques-Bonet T, Eichler EE (2013) Evolution and diversity of copy number variation in the great ape lineage. Genome Res 23:1373–1382
27. Deng C, Cheng CH, Ye H, He X, Chen L (2010) Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc Natl Acad Sci U S A 107:21593–21598
28. Storz JF, Opazo JC, Hoffmann FG (2013) Gene duplication, genome duplication, and the functional diversification of vertebrate globins. Mol Phylogenet Evol 66:469–478

29. Lee JA, Lupski JR (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron 52:103–121
30. Dumas L, Sikela J (2009) DUF1220 domains, cognitive disease, and human brain evolution. Cold Spring Harb Symp Quant Biol 74:375–382
31. Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, Jackson J, Sikela M, Raznahan A, Giedd J, Rapoport J, Nagamani SS, Erez A, Brunetti-Pierri N, Sugalski R, Lupski JR, Fingerlin T, Cheung SW, Sikela JM (2012) DUF1220 domain copy number implicated in human brain size pathology and evolution. Am J Hum Genet 91:1–11
32. Davis JM, Searles VB, Anderson N, Keeney J, Raznahan A, Horwood LJ, Fergusson DM, Kennedy MA, Giedd J, Sikela JM (2015) DUF1220 copy number is linearly associated with increased cognitive function as measured by total IQ and mathematical aptitude scores. Hum Genet 134:67–75
33. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260
34. Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF (2006) Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25
35. Zhang Y, Jiang WK, Gao LZ (2011) Evolution of microRNA genes in Oryza sativa and Arabidopsis thaliana: an update of the inverted duplication model. PLoS One 6:e28073
36. Yuan Z, Sun X, Jiang D, Ding Y, Lu Z, Gong L, Liu H, Xie J (2010) Origin and evolution of a placental-specific microRNA family in the human genome. BMC Evol Biol 10:346
37. Sun J, Zhou M, Mao Z, Li C (2012) Characterization and evolution of microRNA genes derived from repetitive elements and duplication events in plants. PLoS One 7:e34092
38. Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet 36:1282–1290
39. Ha M, Lu J, Tian L, Ramachandran V, Kasschau KD, Chapman EJ, Carrington JC, Chen X, Wang XJ, Chen ZJ (2009) Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc Natl Acad Sci U S A 106:17835–17840
40. Abrouk M, Zhang R, Murat F, Li A, Pont C, Mao L, Salse J (2012) Grass microRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole-genome duplication. Plant Cell 24:1776–1792
41. Li J, Musso G, Zhang Z (2008) Preferential regulation of duplicated genes by microRNAs in mammals. Genome Biol 9:R132
42. D'Antonio M, Ciccarelli FD (2011) Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol 7:e1002029
43. Fernández A, Chen J (2009) Human capacitance to dosage imbalance: coping with inefficient selection. Genome Res 19:2185–2192
44. Felekkis K, Voskarides K, Dweep H, Sticht C, Gretz N, Deltas C (2011) Increased number of microRNA target sites in genes encoded in CNV regions. Evidence for an evolutionary genomic interaction. Mol Biol Evol 28:2421–2424
45. Dweep H, Georgiou GD, Gretz N, Deltas C, Voskarides K, Felekkis K (2013) CNVs-microRNAs interactions demonstrate unique characteristics in the human genome. An interspecies in silico analysis. PLoS One 8:e81204
46. Woodwark C, Bateman A (2011) The characterisation of three types of genes that overlie copy number variable regions. PLoS One 6:e14814
47. Henrichsen CN, Chaignat E, Reymond A (2009) Copy number variants, diseases and gene expression. Hum Mol Genet 18:R1–R8
48. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853

49. Wang RT, Ahn S, Park CC, Khan AH, Lange K, Smith DJ (2011) Effects of genome-wide copy number variation on expression in mammalian cells. BMC Genomics 16:562
50. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, Cremer T, Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer 20:399–407
51. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL (2004) A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 36:299–303
52. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R, Tsang P, Curry B, Baird K, Meltzer PS, Yakhini Z, Bruhn L, Laderman S (2004) Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proc Natl Acad Sci U S A 101:17765–17770
53. Zhao X, Li C, Paez JG, Chin K, Jänne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64:3060–3071
54. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5:16–18
55. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592
56. 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 (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426
57. Bièche I, Olivi M, Champème MH, Vidaud D, Lidereau R, Vidaud M (1998) Novel approach to quantitative polymerase chain reaction using real-time detection: application to the detection of gene amplification in breast cancer. Int J Cancer 78:661–666
58. Ponchel F, Toomes C, Bransfield K, Leong FT, Douglas SH, Field SL, Bell SM, Combaret V, Puisieux A, Mighell AJ, Robinson PA, Inglehearn CF, Isaacs JD, Markham AF (2003) Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. BMC Biotechnol 3:18
59. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30:e57
60. Armour JA, Sismani C, Patsalis PC, Cross G (2000) Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Res 28:605–609
61. Birchler JA, Yao H, Chudalayandi S (2007) Biological consequences of dosage dependent gene regulatory systems. Biochim Biophys Acta 1769(5–6):422–428
62. Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO (2011) Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Res 21(12):2004–2013
63. Reymond A, Henrichsen CN, Harewood L, Merla G (2007) Side effects of genome structural changes. Curr Opin Genet Dev 17(5):381–386
64. Ricard G, Molina J, Chrast J, Gu W, Gheldof N, Pradervand S, Schütz F, Young JI, Lupski JR, Reymond A, Walz K (2010) Phenotypic consequences of copy number variation: insights from Smith-Magenis and Potocki-Lupski syndrome mouse models. PLoS Biol 8(11):e1000543
65. Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B (2009) Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell 138(1):198–208
66. Zweier C, de Jong EK, Zweier M, Orrico A, Ousager LB, Collins AL, Bijlsma EK, Oortveld MA, Ekici AB, Reis A, Schenck A, Rauch A (2009) CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. Am J Hum Genet 85(5):655–666

67. Olzscha H, Schermann SM, Woerner AC, Pinkert S, Hecht MH, Tartaglia GG, Vendruscolo M, Hayer-Hartl M, Hartl FU, Vabulas RM (2011) Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. Cell 144(1):67–78
68. LaSalle JM (2011) A genomic point-of-view on environmental factors influencing the human brain methylome. Epigenetics 6(7):862–869
69. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HY, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stütz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO, 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59–65
70. Yeo RA, Gangestad SW, Liu J, Calhoun VD, Hutchison KE (2011) Rare copy number deletions predict individual variation in intelligence. PLoS One 6(1):e16339 (Pharmacogenetics 2004 14(9):615–626)
71. Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, Bjornsdottir G, Walters GB, Jonsdottir GA, Doyle OM, Tost H, Grimm O, Kristjansdottir S, Snorrason H, Davidsdottir SR, Gudmundsson LJ, Jonsson GF, Stefansdottir B, Helgadottir I, Haraldsson M, Jonsdottir B, Thygesen JH, Schwarz AJ, Didriksen M, Stensbøl TB, Brammer M, Kapur S, Halldorsson JG, Hreidarsson S, Saemundsen E, Sigurdsson E, Stefansson K (2014) CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature 505(7483):361–366
72. Li X, Tan L, Liu X, Lei S, Yang T, Chen X, Zhang F, Fang Y, Guo Y, Zhang L, Yan H, Pan F, Zhang Z, Peng Y, Zhou Q, He L, Zhu X, Cheng J, Zhang L, Liu Y, Tian Q, Deng H (2010) A genome wide association study between copy number variation (CNV) and human height in Chinese population. J Genet Genomics 37(12):779–785
73. Dauber A, Yu Y, Turchin MC, Chiang CW, Meng YA, Demerath EW, Patel SR, Rich SS, Rotter JI, Schreiner PJ, Wilson JG, Shen Y, Wu BL, Hirschhorn JN (2011) Genome-wide association of copy-number variation reveals an association between short stature and the presence of low-frequency genomic deletions. Am J Hum Genet 89(6):751–759
74. Ukkola-Vuoti L, Kanduri C, Oikkonen J, Buck G, Blancher C, Raijas P, Karma K, Lähdesmäki H, Järvelä I (2013) Genome-wide copy number variation analysis in extended families and unrelated individuals characterized for musical aptitude and creativity in music. PLoS One 8(2):e56356
75. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86(5):749–764
76. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43(9):838–846
77. Pollex RL, Hegele RA (2007) Copy number variation in the human genome and its implications for cardiovascular disease. Circulation 115(24):3130–3138
78. Chapman J, Rees E, Harold D, Ivanov D, Gerrish A, Sims R, Hollingworth P, Stretton A, GERAD1 Consortium, Holmans P, Owen MJ, O'Donovan MC, Williams J, Kirov G (2013) A genome-wide study shows a limited contribution of rare copy number variants to Alzheimer's disease risk. Hum Mol Genet 22(4):816–824

79. Gajecka M, Mackay KL, Shaffer LG (2007) Monosomy 1p36 deletion syndrome. Am J Med Genet C Semin Med Genet 145:346–356
80. Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, Porter K, Prigmore E, Krepischi-Santos AC, Varela MC, Koiffmann CP, Lees AJ, Rosenberg C, Firth HV, de Silva R, Carter NP (2006) Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet 38:1032–1037
81. Nevado J, Mergener R, Palomares-Bralo M, Souza KR, Vallespín E, Mena R, Martínez-Glez V, Mori MÁ, Santos F, García-Miñaur S, García-Santiago F, Mansilla E, Fernández L, de Torres ML, Riegel M, Lapunzina P (2014) New microdeletion and microduplication syndromes: a comprehensive review. Genet Mol Biol 37:210–219
82. Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552–564
83. Dixit A, McKee S, Mansour S, Mehta SG, Tanteles GA, Anastasiadou V, Patsalis PC, Martin K, McCullough S, Suri M, Sarkar A (2013) 7q11.23 Microduplication: a recognizable phenotype. Clin Genet 83:155–161
84. Malhotra D, Sebat J (2012) CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell 148(6):1223–1241
85. Peñagarikano O, Abrahams BS, Herman EI, Winden KD, Gdalyahu A, Dong H, Sonnenblick LI, Gruver R, Almajano J, Bragin A, Golshani P, Trachtenberg JT, Peles E, Geschwind DH (2011) Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits. Cell 147(1):235–246
86. Doherty JL, Owen MJ (2014) Genomic insights into the overlap between psychiatric disorders: implications for research and clinical practice. Genome Med 6(4):29
87. Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, van Kessel AG (2010) Germline copy number variation and cancer risk. Curr Opin Genet Dev 20(3):282–289
88. Krepischi AC, Pearson PL, Rosenberg C (2012) Germline copy number variations and cancer predisposition. Future Oncol 8(4):441–450
89. Jin G, Sun J, Liu W, Zhang Z, Chu LW, Kim ST, Sun J, Feng J, Duggan D, Carpten JD, Wiklund F, Grönberg H, Isaacs WB, Zheng SL, Xu J (2011) Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. Carcinogenesis 32(7):1057–1062
90. Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Sun J, Wiklund F, Wiley K, Isaacs SD, Stattin P, Xu J, Duggan D, Carpten JD, Isaacs WB, Grönberg H, Zheng SL, Chang BL (2009) Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. Cancer Res 69(6):2176–2179
91. Yang L, Liu B, Huang B, Deng J, Li H, Yu B, Qiu F, Cheng M, Wang H, Yang R, Yang X, Zhou Y, Lu J (2013) A functional copy number variation in the WWOX gene is associated with lung cancer risk in Chinese. Hum Mol Genet 22(9):1886–1894
92. Huang L, Yu D, Wu C, Zhai K, Jiang G, Cao G, Wang C, Liu Y, Sun M, Li Z, Tan W, Lin D (2012) Copy number variation at 6q13 functions as a long-range regulator and is associated with pancreatic cancer risk. Carcinogenesis 33(1):94–100
93. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463(7283):899–905
94. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW,

Getz G, Meyerson M, Beroukhim R (2013) Pan-cancer patterns of somatic copy number alteration. Nat Genet 45(10):1134–1140

95. Solimini NL, Xu Q, Mermel CH, Liang AC, Schlabach MR, Luo J, Burrows AE, Anselmo AN, Bredemeyer AL, Li MZ, Beroukhim R, Meyerson M, Elledge SJ (2012) Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science 337(6090):104–109
96. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, Shih DJ, Hovestadt V, Zapatka M, Sturm D, Jones DT, Kool M, Remke M, Cavalli FM, Zuyderduyn S, Bader GD, VandenBerg S, Esparza LA, Ryzhova M, Wang W, Wittmann A, Stark S, Sieber L, Seker-Cin H, Linke L, Kratochwil F, Jäger N, Buchhalter I, Imbusch CD, Zipprich G, Raeder B, Schmidt S, Diessl N, Wolf S, Wiemann S, Brors B, Lawerenz C, Eils J, Warnatz HJ, Risch T, Yaspo ML, Weber UD, Bartholomae CC, von Kalle C, Turányi E, Hauser P, Sanden E, Darabi A, Siesjö P, Sterba J, Zitterbart K, Sumerauer D, van Sluis P, Versteeg R, Volckmann R, Koster J, Schuhmann MU, Ebinger M, Grimes HL, Robinson GW, Gajjar A, Mynarek M, von Hoff K, Rutkowski S, Pietsch T, Scheurlen W, Felsberg J, Reifenberger G, Kulozik AE, von Deimling A, Witt O, Eils R, Gilbertson RJ, Korshunov A, Taylor MD, Lichter P, Korbel JO, Wechsler-Reya RJ, Pfister SM (2014) Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511(7510):428–434
97. Liu B, Yang L, Huang B, Cheng M, Wang H, Li Y, Huang D, Zheng J, Li Q, Zhang X, Ji W, Zhou Y, Lu J (2012) A functional copy-number variation in MAPKAPK2 predicts risk and prognosis of lung cancer. Am J Hum Genet 91(2):384–390
98. Nik-Zainal S, Wedge DC, Alexandrov LB, Petljak M, Butler AP, Bolli N, Davies HR, Knappskog S, Martin S, Papaemmanuil E, Ramakrishna M, Shlien A, Simonic I, Xue Y, Tyler-Smith C, Campbell PJ, Stratton MR (2014) Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nat Genet 46(5):487–491
99. Parl FF (2005) Glutathione S-transferase genotypes and cancer risk. Cancer Lett 221(2):123–129
100. Schoedel KA, Hoffmann EB, Rao Y, Sellers EM, Tyndale RF (2004) Ethnic variation in CYP2A6 and association of genetically slow nicotine metabolism and smoking in adult Caucasians. Pharmacogenetics 14(9):615–626
101. Hebbring SJ, Adjei AA, Baer JL, Jenkins GD, Zhang J, Cunningham JM, Schaid DJ, Weinshilboum RM, Thibodeau SN (2007) Human SULT1A1 gene: copy number differences and functional implications. Hum Mol Genet 16(5):463–470
102. Brauch H, Schroth W, Goetz MP, Mürdter TE, Winter S, Ingle JN, Schwab M, Eichelbaum M (2013) Tamoxifen use in postmenopausal breast cancer: CYP2D6 matters. J Clin Oncol 31(2):176–180
103. Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, Xu XH, Yan H, Liu X, Qiu C, Zhu XZ, Chen T, Li M, Zhang H, Zhang L, Drees BM, Hamilton JJ, Papasian CJ, Recker RR, Song XP, Cheng J, Deng HW (2008) Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am J Hum Genet 83(6):663–674
104. Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, Garg S, Henning E, Blackburn H, Loos RJ, Wareham NJ, O'Rahilly S, Hurles ME, Barroso I, Farooqi IS (2013) Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat Genet 45(5):513–517
105. Ensenauer RE, Adeyinka A, Flynn HC, Michels VV, Lindor NM, Dawson DB, Thorland EC, Lorentz CP, Goldstein JL, McDonald MT, Smith WE, Simon-Fayard E, Alexander AA, Kulharya AS, Ketterling RP, Clark RD, Jalal SM (2003) Microduplication 22q11.2, an emerging syndrome: clinical, cytogenetic, and molecular analysis of thirteen patients. Am J Hum Genet 73:1027–1040
106. Wentzel C, Fernström M, Ohrner Y, Annerén G, Thuresson AC (2008) Clinical variability of the 22q11.2 duplication syndrome. Eur J Med Genet 51:501–510

Chapter 7
Mini- and Micro-Satellite Markers in Health, Disease and Evolution

Vasiliki A. Galani, Sofia Markoula, Leandros Lazaros, Paris Ladias, and Ioannis Georgiou

Abbreviations

CNS	Central nervous system
CNVs	Copy number variations
DSBs	Double strand breaks
FSHD	Facioscapulohumeral muscular dystrophy

V.A. Galani, Ph.D.
Department of Anatomy – Histology – Embryology, University of Ioannina,
Ioannina 45110, Epirus, Greece
e-mail: vgalani@cc.uoi.gr

S. Markoula, M.D., Ph.D.
Department of Neurology, University Hospital of Ioannina, Ioannina 45110, Epirus, Greece
e-mail: smarkoula@grads.uoi.gr

L. Lazaros, Ph.D.
Division of Reproductive Genetics, Department of Obstetrics & Gynecology,
University of Ioannina, Ioannina 45110, Epirus, Greece

Laboratory of Medical Genetics and Clinical Embryology, Department of Obstetrics &
Gynecology, University Hospital of Ioannina, Ioannina 45110, Epirus, Greece
e-mail: leandroslazaros@yahoo.com

P. Ladias, B.Sc.
Division of Reproductive Genetics, Department of Obstetrics & Gynecology,
University of Ioannina, Ioannina 45110, Epirus, Greece
e-mail: parisladias@hotmail.com

I. Georgiou, Ph.D. (✉)
Division of Reproductive Genetics, Department of Obstetrics & Gynecology,
University of Ioannina, Ioannina 45110, Epirus, Greece

Division of Medical Genetics and Clinical Embryology, Department of Obstetrics &
Gynecology, University Hospital of Ioannina, Ioannina 45110, Epirus, Greece
e-mail: igeorgio@uoi.gr

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health,
Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_7

G4	G-quadruplex
HD or HTT	Huntington disease
HERVs	Human endogenous retroviruses
HNPCC	Hereditary non-polyposis colorectal cancer
HVR	Hypervariable
LINEs	Long interspersed elements
LP-BER	Long-patch base excision repair
LTRs	Long repetitive sequences
MD	Myotonic dystrophy
MMR	Mismatch repair
MSI	Microsatellite instability
NAHR	Non-allelic homologous recombination
Nt	Nucleotides
ORFs	Open reading frames
PABPs	Poly(A) binding proteins
SBMA	Spinobulbar muscular atrophy or Kennedy disease
SCA	Spinocerebellar ataxia
SCA8	Spinocerebellar ataxia type 8
SINEs	Short interspersed elements
SN-BER	Single nucleotide-base excision repair
SNPs	Single nucleotide polymorphisms
SSRs	Simple sequence repeats or STRs: short tandem repeats
SVAs	SINE-VNTR-ALUs
TEs	Transposable elements
TNR	Trinucleotide repeat
TREDs	Trinucleotide repeat expansion disorders
TRs	Tandem repeats
VNTRs	Variable number tandem repeats

Introduction

DNA repeats are common in eukaryotic and in prokaryotic organisms classified into two main categories. The first category is represented by the interspersed repeats, so-called because they are consisted of repeated units which contain distinct remnants of transposons [1, 2]. Interspersed repeats are quantitatively the most abundant sequences of repeated motifs, occupying approximately 45 % in humans and 33 % in mice, explaining to a certain extent the degree of variation in genome size between different organisms [2]. The second category includes the tandem repeats (TRs) that occur in low complexity DNA when a steady pattern of one, two or many more nucleotides is repeated and the repetitions are not only sequential but also directly adjacent to each other [1]. Because TRs were initially identified as satellite bands in density-gradient centrifugal separations of genomic DNA, the name satellite DNA has been widely used in reference to TRs, although there are marked differences between satellites and TRs. TRs are further stratified in the (a) microsatellites or Short Tandem repeats (STR) the (b) minisatellites or Variable Number Tandem Repeats (VNTRs) and (c) the recently identified megasatellites [3]. Tandem repeats constitute about 8 % of the human genome while only microsatellites comprise approximately 3 % of the sequenced human genome [4]. The sequence CAGCAGCAGCAG, for example, is a short TR, where the repeated unit is the tri-nucleotide CAG sequence, and in this example is repeated four times.

In tandem repetitive DNA sequences in the human genome, in particular micro-satellites and minisatellites, differ from the other repetitive DNAs such as, satellite DNA, telomeres, transposable and retrotransposable sequences and high copy number genes in many ways [5, 6]. In contrast to satellite DNA, as for example alpha satellite DNA, which is found in heterochromatic regions of the chromosomes, minisatellites and microsatellites are generally found in euchromatin [7]. Telomeres, although by definition have a repeat length similar to microsatellites, are unlike to TRs constituted of invariable repetitive sequences of (TTAGGG)n, that are specific to the chromosome tips and responsible for protection from chromosome fusion and rearrangement [8–10]. In contrast also to transposable and retrotransposable sequences, the TRs do not propagate themselves in the genome by a cut and paste or a copy and paste mechanism, but they rather originate from endless cycles of DNA recombination, replication and repair due to their contribution and participation in all these processes [11].

Unlike also to high copy number genes, that are, in a sense, repetitive complex sequences, the TRs are not transcribed autonomously, but they may harbor regulatory elements or extend in exons or introns and interfere with protein synthesis [12].

Tandem Repeats (Micro-, Mini- and Mega-Satellite) Sequence Definitions

TRs are classified according to their size in repeats with units less than nine nucleotides (nt) in length, which are known as microsatellites, or also as simple sequence repeats (SSRs), or short tandem repeats (STRs), and those with units of 10 nt or greater in length that are known as minisatellites [1]. It has been suggested that TRs with enormously long units, greater than 135 nt, comprise a separate class of repeats termed megasatellites [13]. They can be further classified into perfect or imperfect repeats (also called exact or approximate repeats) depending on whether they are precise copies or differ by ≥1 bp due to mismatch mutations, insertions or deletions. Microsatellites are the most prevalent and are almost perfect repeats [3, 14].

Since there was no direct association between the content in DNA repeats of a living organism with its evolutionary history, the repeats were initially regarded as having no biological effect. For this reason they were referred to as non-functional junk or selfish DNA during the first decades in the advent of DNA research [15–17].

Microsatellite DNA

STRs are in general less frequent than SNPS. According to "The 1000 Genomes Project Consortium 2012" there are approximately 700,000 STR loci identified in individual genomes tested at the first phase of the 1000 genome Project. In contrast 3.8 million SNPs were identified in 270 individuals and catalogued by the International HapMap Consortium, by October 2007. Consequently in an approximated frequency of 1SNP per 1000 genome bases, STRs are estimated to less than 1 per 5000 bases.

Microsatellite DNA is formed of much smaller units than any other satellite, with only **1** to 9 bp and they can be found in both coding and non-coding regions. The vast majority of microsatellites have 2 bp repeats the so called dinucleotide repeats, being most frequently (CA)n, followed by (AT)n, (GA)n and (GC)n, with the last combination of repeat being rare [14] or 3 bp repeats consisting either a codon, as for example most frequently (CAG), coding for glutamine, or other triplets without coding properties [14, 18].

There is evidence that the regional variation in microsatellite frequency cannot be explained by their base composition alone [19] and that the density of the microsatellite is almost twice higher at the ends of the arms of chromosomes in the human and mouse genomes [20]. Microsatellites are also frequently found in proximity to interspersed repetitive elements, such as the short interspersed nuclear elements (SINEs) and the long interspersed nuclear elements (LINEs). For example, human ALU repeats are often accompanied by structures resembling to microsatellites at their 3′ ends, that have probably evolved by the insertion of poly(A) tails of reversed transcribed messages that follow retrotransposable elements at their insertion in a new position in the genome [21]. In addition, mononucleotide poly(A) arrays or (A)n and other conformations of A-rich microsatellites predominate at ALU insertions and this is considered as potential evidence for the association of microsatellites with poly(A) tails [14].

More than one million microsatellite loci are dispersed in the human genome. This number includes a significant group of microsatellites interrupted by other sequences and a group that is uninterrupted monomorphic [14]. In addition, mononucleotide repeats, particularly A and T repeats, are encountered in half a million loci, while the number of pentanucleotide repeat loci is only a few thousand and the higher the repeat sequence is, the more rare the microsatellite becomes. It is probable that larger repeats are less common because microsatellites have arisen with the processes of the so called "DNA slippage", "polymerase slippage", or "slipped strand mispairing" [22, 23]. When the new strand is synthesized from the template strand during replication of the microsatellite sequence, it will sometimes pair with another part of the repeat sequence. If the template strand is looped out then a contraction of the microsatellite will result. If the nascent strand loop is integrated, then expansion of the microsatellite will occur [24, 25]. In addition, the recombination events, such as unequal crossing over and gene conversion, can also cause the contractions and expansions of TR sequences. It has been widely discussed and accepted that replication is the most common cause of instability in microsatellites, but recent studies provide evidence that recombination may also cause microsatellite instability [26].

Microsatellite repetitions are found in the order of 10 to 100 times and are therefore the most useful DNA markers for fingerprinting, population, recombination and evolution studies but also for direct and indirect diagnosis in the clinical practice that require polymorphic markers [14, 27–31].

Minisatellite DNA

Minisatellite DNA is consisted of repeated units of 10–400 bp long with an average of about 20 bp. In general the repeat units cover a stretch of approximately 1–5 Kb, with 20–50 repetitions. Specifically, in humans, 90 % of minisatellites sequences are located to subtelomeric regions [32] and are hypervariable stretches of DNAs [33]. The basic repeat unit may vary in length from 10 to >50 nucleotides, with mutation rates ranging from 0.5 to >20 % per generation. They include some of the most variable loci in the human genome and often referred to as **V**ariable **N**umber **T**andem **R**epeats (**VNTR**s), which is probably the most familiar terminology in the majority of the manuscripts. This terminology also highlights the fingerprinting properties of the minisatellites in proximity to genes or in dispersed clusters in euchromatic regions of the chromosomal DNA [27–29]. Minisatellites have been frequently used as genetic markers in linkage analysis studies [34] and also in population studies [35]. They have been considered as regulatory regions of gene expression or as parts of genuine open reading frames [36]. Finally, minisatellites as also microsatellites, are associated with chromosome fragile sites and are found in the vicinity of a number of recurrent translocation breakpoints [37].

Minisatellites have also been reported to act as “hot spots” for homologous recombination [38]. The huge expansions in minisatellite sequences are a result of unequal crossover or of unequal sister chromatid exchange. Thus, these genetic mechanisms probably account for the extreme variability that is often seen between individuals at these loci [39].

Megasatellite DNA

Megasatellite DNA is part of the satellite DNA families, with extremely large tandem repetitive sequences that cannot be classified as minisatellites. They are called either macro- or megasatellite DNA [40–44]. One characteristic example of a megasatellite DNA is the RS447 which is found on human 4p16.1 to consist of 20–100 copies with a 4746 bp unit sequence, containing an open reading frame of 1590 bp encoding an intronless functional deubiquitinating enzyme (USP17) gene [44–46]. The RS447 megasatellite itself contains the functional expression unit of USP17 that is expressed in human cells [46]. While USP17 is ubiquitously expressed in human tissues, presents with a unique expression pattern in the human brain, with its complementary strand transcribed as an antisense transcript that potentially regulates the level of USP17 expression [46]. The high-level expression of USP17 antisense RNA in brain may suppress expression of USP17 and stimulate ubiquitin-dependent protein degradation. Strikingly the copy number of RS447 is hypervariable (HVR) and highly polymorphic between individuals, as well as among mammalian species [44], but the unit sequences of RS447 are extremely homologous within species [47]. In conclusion, the unstable nature of RS447 megasatellite DNA leads to its hypervariability and may contribute to the structural dynamics of this repetitive DNA in the genome.

In addition, the megasatellite repeat DXZ4, localized at Xq23-24, has 50–100 copies of a CpG-rich 3-kb monomer. DXZ4 is found in constitutive heterochromatin on the active X chromosome characterized by a highly structured pattern of H3K9me3 nucleosomes. In contrast, an amount of DXZ4 is found in euchromatin on the inactive X chromosome characterized by the modified histones H3K4me2 and H3K9Ac. It has been put forward, that this megasatellite repeat is implicated in a novel function involving X chromosome inactivation [48].

Tandem Repeats Function in the Genome

Satellites have a spectrum of functions in the genome and were recently recognized as key players in evolution, methylation, regulation and stabilization. Tandem repeats in the coding sequence may cause the generation of toxic or malfunctioning proteins and the non-coding repeats may result in chromosome fragility, in the silencing of the genes in which they are located, in the regulation of transcription and translation, and in the sequestering of proteins that are associated with splicing and cell architecture [49].

Tandem repeats are not uniformly represented in all mammalian genes with an average representation to about 17 % of the genes, but their unique property is that they are often unstable or hypervariable in mammals. Flows in DNA replication and repair are constantly introducing changes in the repeated sequences and in the number of repeat units and result in TR mutations. The frequency of TR mutations depends on the number of repeat units, the repeat consistency and the length of the repeat unit. The mutation rates of TRs are also caused by environmental factors and the higher transcription and replication rates that may lead to increased instability [50].

The longer and the more consistent a repeat region is, the more unstable it becomes and the mutation rates are often 10–100,000 times higher than the average mutation rates in other parts of the genome [25, 51–53]. Tandem repeat polymorphisms are emerging as a third major class of genetic mutation, with counterparts the single nucleotide polymorphisms (SNPs) and the copy number variations (CNVs). Moreover, tandem repeat variability is by definition an important genetic change by itself, but also influences certain changes to concomitant epigenetic marks due to the high instability and the complete reversibility [2].

The most prominent example of instability and versatility of TRs is that most humans have 30 CGG•CGG repeats in the 5′ UTR of their FMR1 gene [54–56], while population studies in Caucasians indicate that ~1 in 246–468 females have 55–200 repeats while ~1 in 3717–8918 males have 200 to >1000 repeats in the FMR1 gene due to instability induced through maternal meiosis [57]. Thus, the TRs in the genomes are polymorphic, with some individuals or families having some tandem repeat regions that are significantly longer than those recorded in the general population. In addition, these larger repeat lengths are not biologically neutral. Indeed, FMR1 alleles with 55–200 CGG•CCG repeats, in the premutation range, are related to neurodegeneration [58] and ovarian insufficiency [59, 60]. While extended alleles with >200 repeats, in the full mutation range, are related to intellectual disability and autistic symptoms [61]. The FMR1 gene is not unique as regards significant repeat length polymorphism and also in the fact that is related to disease pathogenesis. To date, >15 disorders have been identified in humans as the results of the presence of large expansion-prone DNA tandem repeats in their corresponding genes [49].

Many genetic disorders, mostly neurodegenerative and late onset, are caused by dynamic mutations, concerning almost exclusively trinucleotide repeat (TNR) expansions. The aberrant expansion of TNRs and consequently their instability within specific genes in germ line cells and somatic cells are the causes of these genetic disorders [62]. While many repeats are located within non-coding regions, TNRs also occur often within coding and regulatory regions [53, 63]. Several neurodegenerative diseases are associated to variation of specific repeat areas located within coding regions, but variation of repeats located in introns and untranslated regions of genes can also lead to various disorders [62].

In general the repeat expansion diseases that are analyzed separately in the text of this chapter can be stratified into two subclasses:

- Those with TNRs located in an exon coding for polyglutamine tracts
- Those with TNRs in non-coding, regulatory or coding but non-translated sequences [64–66].

Both coding and non-coding TNR repeats can have significant effects on cell processes [62]. TNRs are highly polymorphic and most importantly above a threshold, in average of 30–50 repeat units, the repeats transcend the afforded length and undergo transition to the pathogenic unstable repeat range. TNRs that are located in different parts of the genes, including the 3′ and the 5′ UTRs, the exons and the introns, are related to various disorders with the common feature of repeat instability over a given threshold [67].

Repeat instability can speed up the disease progression, depending on the expansion, the age at onset of the parents and the gender of the carrier parent. The degree of variation of the repeat sequence over time is associated with tissue or cell-type [67]. It has been observed that in most of the CAG/CTG disorders, repeat instability usually affects the brain, except the cerebellum, which presents with reduced repeat instability [67]. Moreover, the fact that the somatic CAG instability is usually greater in the Central Nervous System (CNS) and, more specifically, in the neurons, indicates the implication of inappropriate mismatch repair rather than DNA replication processes as presented in this chapter in the DNA repair and instability section [68].

Variable tandem repeats that are located in promoters of the human genome or coding sequences can act as mediators for rapid phenotypic changes. The frequent contraction or expansion of the repeat tract leads to quantitative and progressive changes in gene expression or function [69]. Thus, TR sequences in the promoters are involved in gene expression variation, suggesting that such sequences confer an important regulatory role due to genetic variation and can accelerate the evolution of gene expression. This is because the inherent instability of tandem repeats in promoters may lead to altered levels of transcription, generated by the polymorphism that allows rapid divergence [69].

On the other hand, while some variable TRs can influence gene expression by modifying the transcription factor binding sites or by spacing between promoter elements, many variable TRs may control promoter activity by modifying the structure of the chromatin itself [2].

In higher vertebrates, TRs also mediate evolvability in organismal morphology and these unstable repeats may give evolutionary flexibility to organ and/or body shape [53].

Microsatellite Functions

The exceptional variability of microsatellite sequences has permitted microsatellites to be utilized in various applications such as genetic mapping, forensic science, conservation, population and evolutionary studies.

Some microsatellites may serve as regulatory elements of gene expression due to their considerable density in promoters, together with their ability to function as structural elements [70, 71]. Promoter related microsatellites are frequently G/C rich, and many are found within or near the 5′ UTR, CpG islands, and G-quadruplex (G4) structures [72].

The highly conserved microsatellites are abundant in the promoter regions of diverse mammalian genes, many of which seem to regulate cell and tissue growth and development [73]. The promoter function is induced by the variation of promoter microsatellites, which leads to a variation of phenotypes. This variation may be beneficial or disease causing [74, 75]. Nevertheless dynamic increases in the repeat lengths of microsatellites found within promoters can at times distort the normal phenotypes and produce abnormal ones [2]. It is well known that the expansion of microsatellites in protein coding or 5′ untranslated regions causes Huntington's disease and fragile-X syndrome respectively [2].

Promoter microsatellites also have the ability to create various DNA secondary structures, with varying consequences in the regulation of gene expression and the chromatin conformation [76]. Indeed, microsatellites with the motif AC/GT can form ZDNA, which is a left-handed spin double helix [77], and microsatellites that consist of the motif AG/CT can form H-DNA a DNA triplex structure [78–80]. Another DNA secondary structure is the G-quadruplex (G4) [81]. The G4 secondary structure plays an important role in gene regulation [82, 83] and can be highly conserved in mammals [83], particularly in promoter regions [82, 84]. These particular DNA structures can either regulate transcription by modulating RNA polymerase activity [85, 86] or by affecting RNA folding when present in 5′ UTR [87, 88].

G/C rich motifs that include the so called CpG dinucleotides are candidate sites for epigenetic modifications, due to the abundance of cytosines. Each one of the total G/C containing microsatellites, with the exception of the unusual mononucleotide motif C/G, includes CpG dinucleotide targets of epigenetic modification [89].

It has been observed that changes in repeat number of these CpG islands in microsatellites would change the number of methylation sites. In addition, changes in microsatellite length may influence the structural potential, which is important because G4 creation appears to limit the methylation at CpG dinucleotides [90]. Thus, the longer the microsatellites are, the higher the probability to have increased structural potential, and may in turn meddle in methylation [91].

Moreover, microsatellites composed of the motif AC/GT can have significant effects, on human phenotypes, when they acquire instability. The promoters of neural development genes include an impressive number of conserved microsatellites [73, 92]. Changes in AC/GT length have been shown to regulate gene expression [93], and the AC/GT length variation is associated with human phenotypes that depend on the regulation of genes harboring AC/GT [72]. It has been also suggested that the microsatellite motif can also influence RNA structure, due to the identification of certain strand-specific biases [92].

A remarkable fact is that the human chromosomes have very low numbers of A repeats and also absence of ACG repeats. This phenomenon could be explained by various arguments:

- These microsatellites play no structural role in the genome and if they occur by chance, they may not be able to maintain during evolution.
- They may offer strong signals for methylation of CG sequences.
- They are probably involved in DNA transcription and replication, possibly through the formation of unusual DNA structures [94].
- Slippage of DNA strands which is essential for the repeats to grow in length [31] may not be possible with some microsatellite sequences and the secondary structures they form.
- The insertion of repeat sequences in the range of 24–40 nucleotides in either nucleosome cores or inter-nucleosomal DNA, affects chromosome structure and cannot be maintained [95]. An observation that underlies the above arguments comes from Figs. 7.1 and 7.2, where our bioinformatic approach clearly demonstrates that the repeat frequencies of certain di- and tri-nucleotides are substantially different throughout the genome and that they cannot be all maintained in equal repeat lengths depending on their nucleotide composition.

Thus, the abundance of microsatellites varies strongly for each sequence motif. Microsatellites with different repeated motifs may be associated with the determination of chromosome structure. Some chromosome structures have also been associated with a higher percentage of AT nucleotides, such as cohesin binding sequence [96] or scaffold attached regions [97]. In conclusion, changes in microsatellite sequence repeat copy number occur mainly in somatic cells [51]. Mutagenesis rate for microsatellite DNA is expected to be in the range of 0.1–0.2 % [51, 98, 99], although a higher mutation rate has also been reported to reach as high as 1.5 % at the human DXS981 locus [100]. In contrast to the above, triplet repeats, that are implicated in several inherited neuromuscular degenerative diseases, can be increased in either male or female germinal cells [101], despite their stability in the majority of the somatic cells with the exception of brain and CNS cells [67].

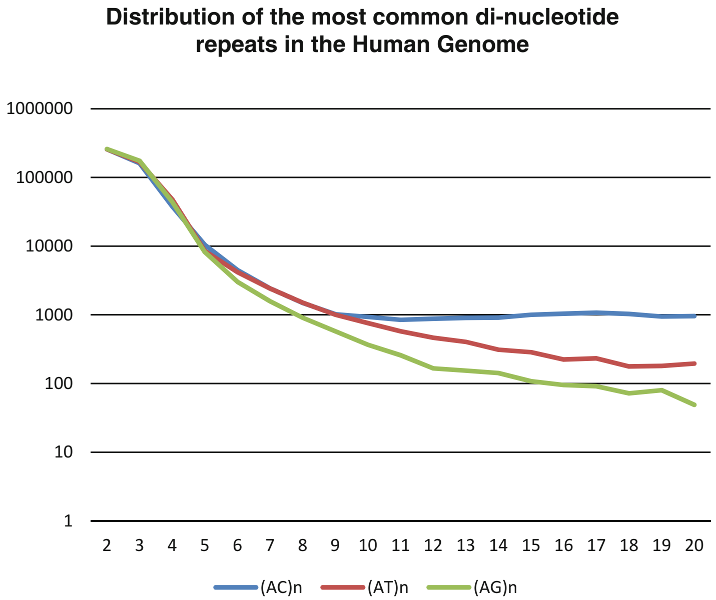


Fig. 7.1 Distribution of most common di-nucleotide repeats in human genome (hg19). The occurrence of dinucleotide sequence lengths are depending on the nucleotide pair variation as demonstrated in this graph concerning the distribution of the most common dinucleotides in the genome. The differences in the distributions support the literature as concerns microsatellite sequence length divergence. The figure clearly shows that the repeat sequence lengths of dinucleotides are different throughout the genome and that microsatellites cannot be all maintained in equal repeat lengths. The vertical axis represents the frequency of the repeats, while the horizontal, the number of the repeats in each particular sequence found in the genome

Minisatellite Functions

Minisatellites are associated with the regulation of genes by various processes. Certain minisatellites are incorporated into open reading frames, some of which may be, polymorphic in the human population [36]. Other minisatellites are binding sites for proteins with a spectrum of functional properties, while minisatellites located in the 5′ region of genes contribute to the regulation of transcription [102]. Certain minisatellites located within introns interfere with splicing, creating multiple splice donor sites [103]. Minisatellites at imprinted loci have been linked to the maintenance of the imprint control [104, 105]. Minisatellites have also been suggested in the literature as intermediate structures in the initiation process of chromosome pairing in eukaryotic genomes [11, 106]. Minisatellites may also compose chromosome fragile sites, identified through induction of replication stress in cell cultures [37]. They have also been found in the locality of a number of repeated translocation breakpoints and in the control recombination site in immunoglobulin heavy chain genes [107].

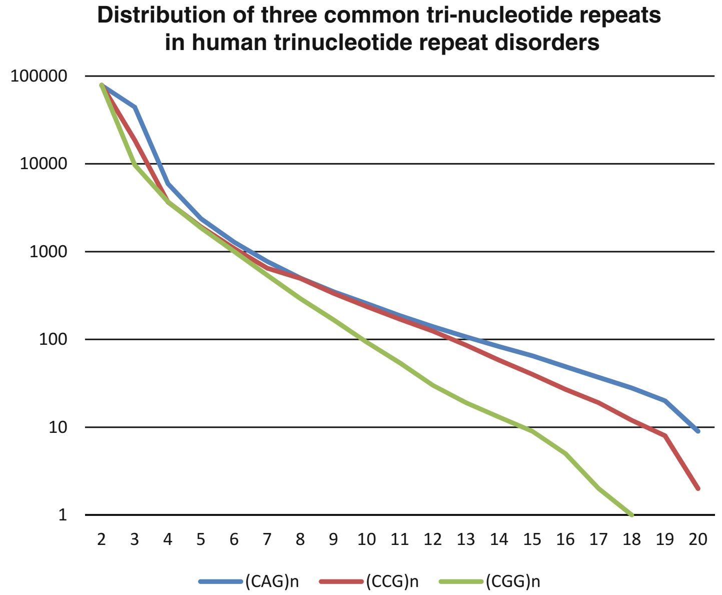


Fig. 7.2 Distribution of most common tri-nucleotide repeats in human genome (hg19). The distribution of three trinucleotide repeats involved in polyglutamine (PolyQ) and non-polyglutamine trinucleotide repeat expansion disorders shows also in this figure a marked microsatellite repeat length divergence. The figure clearly demonstrates that the repeat sequence lengths of certain trinucleotides are substantially different throughout the genome and that microsatellites cannot be all maintained in equal repeat lengths, depending on the three nucleotide combinations (CAG)n in PolyQ disorders, (CCG)nin FRAXE and (CGG)nin FRAXA. The vertical axis represents the frequency of the repeats, while the horizontal, the number of the repeats in each particular sequence found in the genome

In addition, one subcategory of minisatellites includes the highly polymorphic arrays of short tandem repeats with no known function that are used as powerful DNA markers, referred to as variable number tandem repeats (VNTRs). These sequences usually contain 1–5 kb of DNA with repeated units of 15–100 nucleotides [34, 39]. Several minisatellites across the genome share enough sequence homology to be amplified and consequently analyzed by a single primer pair and one probe, yielding efficient DNA fingerprints. Such is a 10–15 bp core sequence of myoglobin minisatellites which includes an invariant core sequence (GGGCAGGANG) among several polymorphic VNTR loci [33, 108].

In conclusion, minisatellites are the most variable loci in the mammalian and human genomes and are therefore excellent molecular markers for studying TR variability and instability. Simple intra-allelic rearrangements are the result of low levels of somatic instability, while complex gene conversions are due to the high frequency of germline instability, which is almost certainly occurring at meiosis. Increased level of instability in microsatellites and minisatellites coincides as a prominent concurrent phenomenon in many human cancer cells [109, 110], atherosclerotic plaques [111] and cells derived from irradiated animals [112, 113].

Telomere Function

Satellites by definition, contain also a particular class of long TRs, located at the tips of all chromosomes which are known as telomeres. Telomeric DNA has 10–15 kb of hexanucleotide repeats (mostly TTAGGG) and is added to the telomeres of all chromosomes by the enzyme telomerase. Such DNA is most certainly functional because it protects the ends of chromosomes from degradation and provides a way for the complete replication of telomeric sequences. It also partially regulates the pairing and the orientation of chromosomes during cell division. Functional aspects that apply to the micro- and mini-satellites, as regards their dynamic behavior, apply also to the telomeres. Telomeric DNA sequences belong, according to the sequence definition given before in this chapter, to microsatellites but some authors regard them as yet another subgroup of minisatellites. The telomeric sequences have 10–15 kb of hexanucleotide repeats, most usually TTAGGG in the human genome, at the tips of the chromosomes. A dedicated enzyme telomerase is occupied solely with ensuring that the complete replication of the chromosome is achieved by adding these sequences perpetually to the telomeres. Telomeres of somatic cells are generally shorter than in germ cells that are designed to live and replicate longer. In humans, it has been suggested that telomeric loss is associated with ageing and tumorigenesis and this is also a landmark being tested in research involving senescence [114].

Megasatellite Functions

The macro- or megasatellite DNA copy number may have an influence on gene expression and the clinical phenotype and may sometimes cause genetic disease [115]. In facioscapulohumeral muscular dystrophy (FSHD), a lower copy number of 3.3-kb tandem repeat D4Z4, is strongly associated with both lower expression levels of adjacent genes and disease severity [42]. The copy number and the methylation status in RS447 DNA may affect both the chromatin structure and the expression of genes in the immediate neighboring region [116, 117].

Moreover, many neurodegenerative diseases, such as Parkinson's disease, are associated with insufficiency of the ubiquitin-proteasome system [118]. Thus, the expression level of the USP17 gene, which may be associated with RS447 copy number and methylation status, may also affect the activity of USP17 deubiquitinating enzyme and cause dysfunction of the ubiquitin-proteasome system, which in turn may alter or even cause the disease [47].

Tandem Repeats Evolutional Drive and Recombination

The scientific community has previously considered regions of tandem repeat DNA as "junk" DNA without any significant mutagenic or selective function in the human genome. However, recent extensive research showed evidence that these DNA fragments have critical functions in the human genome going under natural selection through evolution. Tandem repeats are evolutionary pertinent due to their powerful mutagenic nature. They are mutating in a rate between 10^{-3} and 10^{-6} per cellular division and three phenomena are correlated with TR's evolutionary instability: replication slippage, repair and recombination. During replication, errors occur in result of slip-strand mispairing. Some of them are corrected by exonucleolytic proofreading but many escape this repairing process e.g. by creating secondary structures that escape DNA repairing mechanism. During recombination events, errors occur through unequal crossing-over or by gene conversion that can change TR's repeat number. Interaction between replication slippage, repair and recombination could also lead to mutations as in the case of many neurological diseases, where recombination repair through gene conversion is believed to be responsible for trinucleotide expansions [119].

Strong evidence suggests that tandem repeats are not only randomly distributed along human genome but that their base composition, length and position are conserved and modulated in strict terms through evolution. The establishment of tandem repeats in the ORFs of 17 % of human genes indicates their prevalence over time and their evolutionary impact [2]. An interesting finding by Sawaya et al. [72] made clear that microsatellites can function as regulators of gene expression. This is justified by the high density of microsatellites in promoters which are highly conserved, combined with their ability to function as structural elements. Promoter microsatellites are significantly associated with CpG islands, G4 quadruplexes and untranslated regulatory regions. Variation within these regions could be either evolutionary beneficial to genes that are selectively transcribed or potentially harmful as regards expression of diseases associated genes. Furthermore, Riley and Krieger [120] made an interesting observation for some dinucleotide repeats and their flanking sequences. The findings of their study were informative of the nature and history of tandem repeats and more specifically of the dinucleotide repeats. They found 252 human genes containing $(AC)_{n,}(GT)_{n,}$ $(AG)_{n,}$ or $(CT)_n$ repeats in their UTRs of which 22 had conserved upstream flanking sequences by comparing two different species (Homo sapiens and marsupial species). More interestingly, 18 from these 22 genes were proven to have critical functions in the mammalian embryonic nervous system and an additional one was responsible for kidney's podocyte cells development, cells that have many similarities to neurons. As long as these structures are important for gene transcription, translation, chromatin organization, recombination and DNA replication, natural selection seems to act in a non-random way as regards to TRs size expansions/contractions or base alteration.

Tandem Repeats in Homologous and Non-homologous Recombination

Recombination in all genomes drives evolution and permits meiotic and mitotic changes that accumulate overtime. The gradual accumulation of genetic alterations over a long period of time in reproductive or somatic cells would result in a large-scale reformulation and reconstruction of the human genome. Recombination is clustered in regions of the genome called "hotspots" that undergo recombination in higher rates than average and contribute to genome diversity. Hotspots are distributed in many locations of the human genome as a result of the non-random DSBs (double strand breaks) that initiate recombination events. Regions enriched for these hotspots include telomeres with their flanking sequences, as well as other regions containing repetitive sequences and sequences enriched in GC content. The major conserved regulator protein of recombination hotspots is PR domain zinc finger protein 9, encoded by the PRDM9 gene. PRDM9 is a meiosis-specific histone methyltransferase with a tandem repeat zinc-finger domain encoded by a minisatellite sequence [121]. This specific protein is considered to trigger chromatin remodeling via histone 3 lysine 4 methylation allowing SPO11 to create DSBs so as to initiate recombination [122]. The bizarre in this case is that PRDM9's binding affinity depends on variations in the minisatellite itself enhancing an interesting but yet speculative theory that PRDM9 is driving its own evolution [123]. The fact that PRDM9 regulates about 40 % of human hotspots [124] combined with the discovery of a new hotspot in human genome that is essentially a tandem repeat of a hypervariable minisatellite, with a recombinative potential up to 13.5 times over average [38], declares that some minisatellites are major regulators of human hotspots and recombination.

On the other hand, several reports declare an association between microsatellite polymorphisms recombination rates in the human genome but no one has answered yet the question whether the recombination is mutagenic to microsatellites or microsatellites act as recombination signals. Although microsatellites were considered as genetic markers that have no contribution in phenotypic variations in the past, an interesting finding though by Biet et al. [125] showed significant association of dinucleotide repeats with recombination enzymes in yeast, bacteria and humans. More specific, a common denominator affecting the binding of the recombination enzymes scRad51, hsRad51, ecRecA to oligonucleotide dinucleotide repeats (GT, CA, CT, GA, GC, AT) is the formation of secondary structures such as Z-DNA in ssDNA that is conserved through evolution. As regards to their binding preference, oligonucleotides with CT, GT, CA repeats have stronger binding affinity than GA, AT, CG repeats. As one would expect, di-nucleotide repeats according to Biet, are positively correlated with recombination rates. But, according to Guo et al. [126], the correlation between dinucleotide repeats and recombination rates seem not to be significant compared to motifs of other lengths such us tri-, tetra- or mono-nucleotide repeats. More specifically, dinucleotide microsatellites with different motifs consisting of either A or T lead to increased recombination rates compared to dinucleotide microsatellites with motifs consisting of both A & T together, lead to decreased recombination rates in the human genome. Tri-nucleotide repeats on the other hand show the strongest direct effect on recombination rates in the human genome. That can be explained by the fact that gene densities are remarkably linked with higher recombination rates [127] and are abundant of trinucleotide repeats [128]. But, the question remains whether do microsatellites themselves act as recombination signals. It is commonly known that microsatellites are associated with meiotic hotspots in the MHC loci, as well as in regions with high recombination rates on human chromosome 22 [129], but this is a more complex and multifactorial phenomenon. According to Myers et al. [130], recombination is clustered in "hotspot", regions that have much more recombination potential than the genome average, where certain microsatellite motifs are overrepresented. So the question is why these certain motifs are overrepresented in hotspots. In addition it is unclear whether the majority of hotspots contain microsatellites. An interesting report showed that the major factor which results in converting microsatellite motifs into recombination signals is their base composition affecting the stability of the DNA helix [131]. Furthermore, an interesting report showed flanking sequences have undergone the same evolutionary drive as microsatellites themselves and act as regulators regarding their microsatellite's sequence variation and their motif length [132]. The theory that certain sequence motifs, neighboring to or inside the microsatellites, drive microsatellite and genome diversity through evolution, gains more and more support [132].

Tandem Repeats in Replicative Recombination

Almost 45 % of the human genome is composed of DNA segments related to transposable elements (TEs) which represent truncated of full length transposon sequences within genomes commonly known also as interspersed repeats. Approximately 90 % of these DNA sequences are related to retrotransposable elements, but only a minority of those is active and can exhibit new transposition events. TEs are divided in two major categories: (1) the DNA transposons and (2) the retrotransposons. DNA transposons represent nearly 3 % of the human genome and are characterized by their ability to "cut" themselves from a genomic locus and migrate to another genomic locus (cut and paste). Although DNA transposons lost their ability to propagate themselves in the human genome during evolution, they had highly active role before 38 million years. On the other hand, the retrotransposons are characterized by their ability to copy themselves through RNA to DNA reverse transcription (copy and paste). The outcome of this procedure is the increase of retrotransposon copies in the human genome over time. The retrotransposons are subdivided in two categories, focusing on the presence or absence of long repetitive sequences (LTRs). Human endogenous retroviruses (HERVs) are the so called LTR elements constituting approximately 9 % of the human genome. Most of the HERVs were introduced in the genome before 25 million years and their activity is quite limited to humans. In contrast, the majority of TEs in the human genome is the outcome of the activity of non-LTR elements including LINEs, ALUs, and SVAs which represent nearly 1/3 of the human genome [133]. Data indicate that certain LTR and non-LTR elements have still an active role in human genome, so it is of outmost importance to define the relationship between tandem and interspersed repeats.

Increasing evidence show a strong association between non-LTR elements and microsatellite sequences. More specific, microsatellite sequences can arise during retrotransposition of non-LTR elements, mainly LINE1 and ALU subfamilies, into the genome. Two specific studies showed that 36 % of mono-, di-, tri-, tetranucleotide microsatellites were "born" during the integration of LINEs and ALUs into three primate genomes (human, chimpanzee, orangutan) and that as high as 25 % of microsatellite "births" and 24 % of microsatellite "deaths" occur within LINE1 and ALU sequences during retrotransposition [134]. A common feature of these two major families of non-LTR elements is a poly(A) tail at their 3′ end of their sequence which gives them the unique ability to mobilize [135]. This poly(A) tail is a crucial component of the retrotransposition process and its length strongly influences the dynamics of these major retroelement families [136]. In LINES, a poly(A) RNA tail is normally formed during the transcription process as a 3′ end of LINE mRNA. Contrary to LINEs, only active ALUs have the poly(A) RNA tail which is recognized by LINE ORF1 protein and therefore undergo LINE1 mediated retrotransposition [137]. Poly(A)RNA tails are the binding sites of PABPs (Poly(A) binding protein) creating a critical ribonucleoprotein complex between LINE1 proteins and LINE/ALU RNAs. Several reports demonstrated that the length of these poly(A) tails is positively correlated with LINEs and ALUs ability to integrate into genomes [138, 139]. Not only the overrepresentation of poly(A) repeats in human genome is justified by poly(A) tails of both LINEs and ALUs [128] but also di- and tetra-microsatellite repeats derived from poly(A) tails [21, 140]. Furthermore, the abundance of non-LTR elements in human genome made the perfect environment for the conversion of CAA to NAA, a mechanism responsible for the generation of a large number of A-rich trinucleotide repeats in human genome. More than 60 % of these repeats are located within ALU poly(A) tail sequences [141].

Another interesting non-autonomous non-LTR retrotransposon, that is currently active via LINE1 retrotransposition machinery, is SVA (SINE-VNTR-ALU). SVAs, which are the newest class of retroelements, also members of the family of non-autonomous retrotransposons, have a special feature that makes them unique. Being the only active composite retrotransposons, each component of an SVA is derived either from a retrotransposon or from a repeat sequence, characterizing SVAs as a repeat of repeats. A typical SVA is consisted from the 5′ prime end to the 3′ prime, of a hexameric CCCTCT repeat, followed by a sequence homologous to antisense ALU sequences, a variable number of tandem minisatellite repeats (VNTR domain), and finally a sequence homologous to HERVK-10 endogenous retrovirus, followed by a polyadenylation (polyA) sequence [142]. SVA's core sequence is a VNTR (Variable Number Tandem Repeat) but its functional role is not yet understood, although its length has increased through evolution. SVAs are highly polymorphic, and therefore each repeat domain is capable of misaligning with another locus containing similar SVA element or another repeat sequence, leading to NAHR (non-homologous recombination) [143]. In conclusion, older and younger retroelements are associated with satellite DNA sequences and have a potentially synergistic natural history in the genome. In particular the youngest human specific retroelements, SVAs, have incorporated in their complex genome sequence, parts from microsatellites, minisatellites and other older retroelements.

Repeat Instability Through Meiosis and Repair

Since the discovery of triplet repeat expansions molecular and structural studies explored the hypothesis, that intermediate unstable secondary structures were the culprits of unstable expansions [144–146]. The length of the repeat and the probability of a given sequence to form a secondary structure determine, to a certain point, the likelihood of repeat expansions in polyglutamine or non-polyglutamine disorders [147–149].

Although expansions can arise in both germ cells and somatic cells, anticipation in inheritance of trinucleotide repeat disorders is associated with repeat instability in the germ cells [150]. Furthermore expansions are found in dividing as well as in non-dividing cells meaning that dynamic mutations occur either during DNA replication or during DNA repair [145].

Triplet Repeat Instability at Meiosis

Male meiosis is rather more permissive to trinucleotide expansion studies than female meiosis and oocyte studies. This is due to the very high numbers of spermatozoa in the ejaculate and also to the subsequent high rates of replication cycles during spermatogenesis. A detailed study to identify the particular gametogenesis stage during male germ cell development, at which instability takes place, would elucidate the mechanisms of trinucleotide expansions. Therefore gene expression involved in replication, repair and recombination may answer questions related to the effects of male germ cell mitosis, meiosis or maturation on instability [26]. In the pre-meiotic stage, expansions can occur during mitotic replication and repair, while in the post-meiotic stage expansions can be influenced by meiotic recombinations. Studies of the mismatch repair protein genes in both humans and mice have shown that both MSH2 and MSH3 have expression variability with the first being highly expressed in mitotic spermatogonia, while MSH3 increases expression to peak values at the meiotic spermatocyte stage [151]. Consequently these results may indicate that abnormal repeat expansions may arise as early as at the stage of replicating spermatogonia. In Huntington Disease (HD), which is the most frequent disorder in the group of polyglutamine (CAG)n repeat expansion disorders, elegant studies using modern technology have reported intriguing results. Expansion products were found in almost equal rates in both pre- and post-meiotic cells but the proportion of large expansions in post-meiotic spermatids and spermatozoa was significantly higher [152]. The above indicate that male germ cells expansion in HD can occur before and potentially during or after meiosis [150].

Repeat instability mechanisms during gametogenesis in HD may have similarities to other repeat expansion loci causing instability disorders, but have also marked differences that underline dissimilarities in molecular pathology and genetic counseling. Studies of families with similar (but not CAG) repeat expansion disorders, as is Myotonic Dystrophy (DM1) and Spinocerebellar ataxia type 8 (SCA8), caused by the unstable (CTG)n, have shown repeat expansion bias from the maternal side in contrast to the paternal. In addition, FMR1 gene (CGG)n repeat expansions are invariably maternal and arise in the oocyte, as mothers rather than fathers are responsible for the unstable transmission. FMR1 paternal instability transmission does not take place as (CGG)n stretches contract in FRAXA males to premutation range lengths [153]. Contractions or deletions of the expanded alleles may also be identified in male SCA8 patients in analogy to FRAXA patients [154].

As regards transmission in male HD patients who have a higher risk for sperm cell HTT gene DNA expansions, there could be measures to reduce the (CAG)n instability anticipation. Pearson suggests early pubertal semen storage, to be used for future assisted reproduction, as he postulates from the literature that at the onset of puberty the late onset HD carriers have lower chances to harbor expansions compared to their semen samples later in life [150].

Triplet Instability in Somatic Tissues

The findings in reproductive tissues and germ cells are not necessarily analogous to those of somatic tissues, which undergo completely unrelated to germ cells processes. Differences exist even between tissues of patients or between different trinucleotide repeat expansion disorders. For example different somatic tissue cells from 12 HD patients were found to display repeat mosaicism as in their germ cells [155]. Surprisingly the greatest level of polyglutamine coding (CAG)n mosaicism in HD was detected in both brain and sperm. On top of that finding, affected brain regions by obvious neuropathological lesions were having the highest repeat mosaicism [155].

In a dual experimental approach, both DM patients' fibroblasts were extracted and cultured at progressing ages and DM transgenic mice tissues were studied. The results were intriguing due to the fact that replication progression was associated with the expansions and may affect tissue and age specific repeat instability [156].

In conclusion although germ cells and somatic cells may poses unrelated replication processes, repeat instability is a tissue specific phenomenon that is probably observed predominantly in affected tissues and the germ cells which are known to have varying replication rates but also high plasticity of their DNA [157, 158].

Triplet Repeat Instability Due to Defects in DNA Repair

As presented above CAG instability is more prominent from all somatic cells in the brain neurons of the CNS, implicating mechanisms independent of the replication process [68]. Sound hypothesis and adequate experimentation have suggested and consequently shown that triplet repeats form different secondary DNA structures, depending on their sequence [159]. The secondary DNA specific conformation predetermines stability or instability, explaining why CAG repeats forming predominantly random coils are unstable, while CTGs which tend to form hairpins are stable [160]. Trinucleotide repeat length can also influence stability by generating multiple conformations, raising thus the complexity of the DNA structures [161]. Transcription of the repeats leads to the formation of hybrid DNA–RNA structures in addition to the DNA–DNA structures that may arise in nuclear chromatin and affect the stability of the initial conformation of the repeat secondary structure [162]. Finally, other important nuclear processes may influence the primary and secondary repeat structures, such as methylation, chromatin remodeling and transcription and have an important effect on stability [146].

In mice it has been shown that repeat instability is reduced as a direct consequence of mismatch repair genes inactivation [163, 164]. The same is true for the long-patch base excision repair and the single nucleotide excision repair systems [165, 166].

Base excision repair (BER) is a dedicated DNA repair pathway occupied with the elimination of DNA base flows. Oxidative damage due to 8-oxoguanine (8-oxoG) is the most common DNA lesion [167], that BER is destined to correct by a series of enzymes. The enzymes remove the unmatched DNA base (glycosylase), cleave the abasic site, (Ape1 endonuclease) and repair the either single base defect or the multibase stretch defect (DNA polymerase β-Polβ and flap endonuclease 1-Fen1) [62, 168].

BER protein significance in TNR instability has been studied in repeat instability models of yeast and mice suggesting that replication errors together with BER may contribute to instability. Although there is no clear view which of the two, replication or repair, is more crucial for instability, the disruption of Lig1 and PCNA interaction increased instability due to replication errors in yeast [169]. Lig1 protein deficiency on the other hand did not result in instability, in contrast to Lig1 overexpression that led to repair dependent instability [170]. Similarly, in Lig1 mutant DM mice, a maternal bias with shift to increased contractions and decreased expansions was found, indicating the interplay between replication and repair in instability [171]. Furthermore, in HD mice with *Neil* deficiency somatic and germline instability was reduced, including the brain tissue. Neil1 is a DNA glycosylase acting on pyrimidine-derived lesions, that can also remove duplex and single-strand DNA lesions and 8-oxoG lesions in both somatic and germ line tissues [172]. Finally, in a fragile X premutation model the DNA oxidizing molecule, potassium bromate, exacerbated repeat instability in the germline driving the gene towards the repeat expansion, indicating repair of oxidative DNA lesions as related to the instability of (CGG)n repeats [173].

Age increases oxidative damage on DNA in certain tissues related to repeat instability, as in the germ cells and the brain. Also accessibility of repair proteins to hairpin secondary conformations of (CAG)n and (CTG)n is potentially reduced due to inappropriate recognition [174]. On top, hairpin conformations of (CAG)n and (CTG)n repeats are considered as hot spots for DNA oxidative damage [175]. Consequently, reduced accessibility and increased susceptibility of repeat secondary structures to DNA damage may facilitate accumulation of oxidative lesions at (CAG)nor (CTG)n repeats or even deteriorate to generate unstable expansions [62].

Satellite Repeats in Human Disorders

Satellite repeats, either mini-or micro-satellites with short or long repeated sequences of DNA, are usually prone to acquire repeat instabilities due to errors that occur during the repair process that follows replication. Replication errors, that affect repeat instability, may happen in germ cells at meiosis or in somatic cells at mitosis with defects in recognition and repair of replication errors.

The two major types of satellite repeat involvement in human inherited or acquired disorders are given below:

- Micro- and mini-satellite instability disorders that relate to impaired DNA mismatch repair during replication in somatic cells. Instability disorders are present in cancer cells as the result of the nuclear proofreading machinery failure to correct mismatches that spontaneously arise in fast replicating nuclei. Instability is present in coding, as well as in non-coding microsatellites with potential functional oncogenic effects, either on the chromatin conformation or on the gene expression. Such failures exist (1) due to predisposition owing to inherited defects of the mismatch repair genes (2) due to acquired inability of the cell nucleus to sustain high fidelity replication of the DNA and correct distribution of the genetic material to the daughter cells. Defects in recognition and repair of replication errors are most frequently found in tumors with inactivating mutations on mismatch repair (MMR) genes. Such inactivating mutations are found in disorders inherited from the germ line, such as in human non-polyposis colorectal cancers or in sporadic colorectal, endometrial and gastric cancers. Inactivating MMR gene alterations result in instability of both mini- and micro-satellite repeats in cancer cells and can be used as diagnostic or prognostic markers for these types of cancer.
- Repeat instabilities at meiosis are either contractions or expansions, but the later are those related to human disorders known as the trinucleotide repeat expansion disorders. Triplet repeat expansion disorders, which are common human genetic defects, are characterized by abnormal expansion of a triplet repeat stretch adjacent to a functional gene during meiosis, resulting in an abnormal repeat number interfering with expression or regulation of this gene. The most common repeat forms in expansion disorders are the trinucleotides CAG, CGG, CTG and GAA.

Micro- and Mini-Satellite Instability Disorders

Predisposing germ line mutations, responsible for recognition and repair failures in hereditary non-polyposis colorectal cancer (HNPCC), are more frequently those affecting the MLH1 and MSH2 genes. Both genes belong to classes of mismatch repair homologous genes present in prokaryotes and eukaryotes.

More than 30 different genes have been found to harbor mutations at coding repeat sequences in sporadic cancers with microsatellite instability [176]. Among them are genes directly related to mismatch DNA repair as MLH3, MSH3, MSH6 and PMS2.

Almost all the above mismatch repair genes (MLH1, MSH2, MSH3, MSH6 and PMS2) have been tested for microsatellite mutational biases in different species (yeast, mouse and human) and have been found to favor either deletions or insertions [177].

Cancers may also destabilize minisatellites in patients with MMR gene inherited defects manifested as colorectal cancers, specifically in those patients known to have microsatellite instability (MSI). Furthermore there is evidence that DNA instability may be detected in minisatellites even in cases with a MMR gene mutation, where MSI has not been observed [178].

Repeat Expansion or Trinucleotide Repeat Instability Disorders

Trinucleotide repeat expansion disorders (TREDs) are a defined group of genetic disorders affecting the neural and the neuromuscular system. In a total of more than 15 disorders, (CAG)n and (CTG)n repeat expansions are the most frequent, encountered in the majority of these disorders. The repeats causing the TREDs are polymorphic in the general population with a normal higher repeat range of approximately 30–50 repeats. Loss of function or gain of function is related to the specific gene location of the expansion which is either at the 5′ or 3′ UTR or within introns or exons. Unstable repeats are pathogenic and result in dynamic mutations that can expand further in the germline and somatic tissues and also from one generation to the next [179]. The deterioration of the TREDs from a normal or permutation status to a full mutation is described in medical genetics as anticipation, meaning the clinical presentation of the early onset disorder in the future generations with more pronounced symptoms [26].

There are many important features in each particular TRED, of which the noteworthy are: (1) expansions are more frequent than contractions, (2) germ cells may have the same pathogenic expansions as selective tissue cells, (3) expansions tend to be increased in the affected tissue cells and in particular to the brain cells, (4) repeat sequence tracts in affected somatic tissues seem to increase with age [157, 180].

At least five different trinucleotide sequences are involved in repeat expansion disorders, CGG, CCG, CAG, GAA and CTG, with a frequent range from over 21 tandem repetitions to more than 250 and an excessive range in some particular disorders, as in the FRAX syndrome, with more than 10^3 repeats in full mutations. Maternal meiosis is the common cause of triplet repeat expansions, but contractions may also occur [181].

Various genes in mammalian species and in humans contain repetitive triplets CAG or CAA coding for glutamine. The universal code for glutamine (CAG) is found in more than half of the repeat expansion disorders in humans. Due to the translation of these expansions in the affected cells a deleterious polyglutamine tract is formed leading to polyglutamine or alternatively termed polyQ disorders. The cells may not effectively get rid of proteins with excessively long polyglutamine tracts that eventually can accumulate in nerve cells over time and damage their cytoplasm. For this reason, the stratification of the repeat expansion disorders is facilitated by the simple discrimination between polyglutamine and non-polyglutamine disorders.

Polyglutamine Disorders (CAG)N

There is a significant association of the polyglutamine tract length with a tendency for early onset appearance of the polyglutamine (polyQ) disorders. The most frequent polyglutamine disorder is Huntington chorea, followed by SBMA (Spinobulbar Muscular Atrophy or Kennedy disease) and the Spinocerebellar ataxias (SCA1, SCA2, SCA3, SCA6, SCA7 & SCA17) (Table 7.1).

Non-polyglutamine Disorders

The most frequent disorders in this group are Fragile X syndrome (FRAXA), Myotonic Dystrophy (DM) and Friedreich's ataxia (FRDA), followed by other less frequent disorders, as FRAXE (Fragile XE mental retardation) and Spinocerebellar ataxias (SCA8 & SCA12) (Table 7.2).

Fragile X syndrome (FRAXA) is the most frequent and prominent disorder among the non-polyglutamine disorders, owing its name to the association of a fragile site on the X chromosome, with mental retardation in karyotyped mentally retarded patients. Although the triplet CGG in fragile X syndrome gene FMR1 codes for arginine, the gene expansion is located at the 5′ prime untranslated region of the gene and therefore affects directly on the regulation of the protein FMRP expressed. The long stretches of >55 repeats of CGG and particularly those of full mutations affect mostly the methylation of this region creating more CpG island targets for methylation. Abnormal methylation interacts with all aspects of the molecular function of the FMR1 gene including regulation, expression and replication.

Contribution of Microsatellite and Minisatellite DNAs to Medical Genetics

Microsatellite and minisatellite DNA variations have played a crucial role in the development of medical genetics. Their use in linkage analysis studies revealed the association of various genes with hundreds of hereditary diseases, helped in the chromosomal abnormality screening and in the mapping of genetic loci susceptible to tumors. Many micro- and mini-satellites found to be functional promoter polymorphisms are acting as common genetic risk factors for disorders, like diabetes, or pathogenic mutations (Huntington Disease, Fragile X Syndrome, Myotonic Dystrophy, Spinocerebellar Ataxias). Finally, microsatellites and minisatellites, constituting a powerful tool for the genome identification, the parentage confirmation and the sexual assault examinations, favored significantly the forensic science. The contribution of these DNA variations in medical genetics is described in the following pages.

Table 7.1 Polyglutamine disorders (PolyQ)

Disorder	Gene/product category	Chromosome	(CAG)n normal range	(CAG)n expanded range	Inheritance	GCid
Huntington Chorea (**HD**)	HTT/Protein binding	4p16.3	6–35	36–250	**AD Late Onset**	**GC04P003076**
Spinobulbar Muscular Atrophy (**SBMA**)	AR/DNA binding	Xq12	9–36	38–62	**X-Linked**	**GC0XP066763**
Spinocerebellar Ataxia, Type 1 (**SCA1**)	SCA1/Protein binding	6p23	6–35	49–88	**AD**	**GC06M016299**
Spinocerebellar Ataxia Type 2 (**SCA2**)	SCA2/RNA binding	12q24.1	14–32	33–77	**AD**	**GC12M111890**
Spinocerebellar Ataxia Type 3 (**SCA3**)	SCA3/Protein binding	14q21	12–40	55–86	**AD**	**GC14M092524**
Spinocerebellar Ataxia Type 6 (**SCA6**)	SCA6/Protein binding	19p13	4–18	21–30	**AD**	**GC19M013317**
Spinocerebellar Ataxia Type 7 (**SCA7**)	SCA7/Protein binding	3p21.1	7–17	38–120	**AD**	**GC03P063825**
Spinocerebellar Ataxia Type 17 (**SCA17**)	SCA17/DNA binding	6q27	25–42	47–63	**AD**	**GC06P170863**
Dentatorubropallidoluysian atrophy (**DRPLA**)	ATN1/Protein binding	12p13.31	6–35	48–93	**AD**	**GC12P007033**

1. *AD* Autosomal Dominant, *AR* Autosomal Recessive, *GCid* GeneCards Identifiers
2. Data collected from www.GeneCards.org

Table 7.2 Non-polyglutamine (non-PolyQ) disorders

Disorder	Gene/product category	Chromosome	Repeat normal range	Expanded repeat range	Triplet gene position	Inheritance	GCid
Fragile X Syndrome (**FRAXA**)	FMR1/RNA binding	Xq27.3	6–53	55–200 Premutation >230 Full Mutation	**CGG 5′ UTR**	**X-Linked**	**GC0XP146993**
Myotonic Dystrophy (**DM**)	DMPK/Protein binding	19q13.3	5–37	>50	**CTG Non-coding**	**AD**	**GC19M046272**
Friedreich's Ataxia (**FRDA**)	FXN/Protein binding	9q21.11	7–34	>100	**GAA Intron**	**AR**	**GC09P071650**
Fragile XE Mental Retardation (**FRAXE**)	FMR2/RNA binding	Xq28	6–35	>200	**CCG 5′ UTR**	**X-Linked**	**GC0XP147582**
Spinocerebellar Ataxia Type 8 (**SCA8**)	SCA8/Protein binding	13q21	16–37	110–250	**CTG is Transcribed to mRNA as CUG repeat with toxicity**	**AD**, **Reduced Penetrance**	**GC13U900338**

1. *AD* Autosomal Dominant, *AR* Autosomal Recessive, *GCid* GeneCards Identifiers
2. Data collected from www.GeneCards.org

Microsatellite DNAs and Medical Genetics

Microsatellite markers constitute powerful tools for testing genetic loci associated with common diseases. For example, Huntington's disease which as described previously, is caused by an expansion of a CAG repeat encoding a polyglutamine stretch in the huntingtin protein, the size of which is inversely associated with the age at onset of Huntington's disease [182]. Similarly, Fragile X syndrome, caused by an expansion of a CGG repeat in the fragile X gene, exceeds a threshold length at which the protein produced by the gene cannot be detected [183]. STRs have also been used for the prenatal diagnosis of Down syndrome using amniotic cells [184], the detection of hemophilia A and B [185, 186], the diagnosis of prostate adenocarcinoma in transrectal prostate biopsy specimens [187], the detection of Duchenne muscular dystrophy [188] and the duplication screening in Charcot–Marie-Tooth patients [189].

The requirement of very small amount of DNA, pure or even degraded to some extent, for the analysis of microsatellite markers as well as their adaption to high-throughput systems favored their use in linkage analysis genetics. With the development of the second generation genetic linkage maps of human chromosomes, the responsible genes and the respective mutations for over 2000 genetic disorders were identified. Microsatellite markers have played an important role in the linkage analysis of recessive diseases with very low incidence, in which only a very small number of patients can be collected [190]. In cases of patients, whose ancestors lived in the same area for a long period, only few patients are sufficient to map a gene associated with a genetic disease. The responsible genes for Fukuyama-type congenital muscular dystrophy, an autosomal recessive, severe muscular dystrophy associated with brain anomalies [191], and for the Gelatinous drop-like corneal dystrophy, an autosomal recessive disorder characterized by severe corneal amyloidosis leading to blindness [192], were isolated using microsatellite markers. Finally, the analysis of STRs in the Xq11–Xq13 interval can provide a simple and rapid scan of the genetic loci associated with prostate carcinoma predisposition in large populations [193].

STRs, due to their high heterozygosity, are very useful in paternity and forensic testing. The genotyping of autosomal, Y-chromosomal and mitochondrial STRs has increased the ability to solve problems relevant to paternity or relevant to relatedness on the maternal or the paternal lineage [194–196]. On the other hand, forensic genotyping, providing reliable evidence for sentencing the offenders and exonerating the innocent suspects, has great impact on the society. The use of STR analyses offered a doubtless identification of an individual implicated in violent crimes such as murders or sexual assaults and the identification of remains of missing persons or victims of mass disasters [197–199].

Furthermore, the analysis of autosomal, Y-chromosomal and mitochondrial microsatellites has been widely used for the identification of "founder effect" phenomena [200, 201], which are caused when a small group of individuals cuts out from a larger population and establishes a new population. These populations are usually characterized by increased frequencies of certain genetic diseases, such as Tay–Sachs disease in Ashkenazi Jews [202] and asthma in Hutterites [203].

Consequently, the study of founder populations is very useful for identifying which genes are involved in a genetic disease and which is the genetic profile of the disorder. STRs constitute a valuable tool for the above analysis offering a fine-mapping of the disease genes.

DNA typing is necessary for the engraftment documentation after allogeneic bone marrow transplantation or allogeneic peripheral blood stem cell transplantation [204]. The serial chimerism analysis of STRs offers a reliable and simple screening for the detection of relapse and the identification of patients with progressive mixed chimerism [205]. In case of chronic myeloid leukemia, although the common STR analysis is an appropriate screening test for patients in the post-transplant setting, the use of leukemia STR specific sensitive molecular assay is necessary in patients exhibiting mixed hematopoietic chimerism [206]. Finally, STR analysis is very useful in leukemia patients with graft failure after a bone marrow transplant and in need of a second transplant [207].

Minisatellite DNAs and Medical Genetics

The minisatellite DNAs are highly polymorphic sequences with high heterozygosity in given populations [34], frequently encountered in the literature as variable number of tandem repeats (VNTRs).

Alec Jeffreys, who discovered the technique of genetic fingerprinting in 1984, showed that minisatellites constitute the most variable markers of human genome, exhibiting this variation in the numbers of repeat units or “stutters”. The hypervariation of VNTRs has been utilized for the discrimination of individual genomes in criminal forensic studies and for parentage testing [208]. The use of VNTRs in such cases offered a reliable ‘DNA fingerprinting’, unique for any individual, and a powerful tool for criminal justice [209]. However, allele databanks for every population are essential so as to estimate the probability of a particular allele combination [210]. The analysis of highly variable VNTRs led to a trustworthy paternity testing and eliminated the non-paternity, which constituted a common source of bias in the estimation of mutation rates when they were obtained from family data with discordance of parental and offspring genotypes [211].

The use of VNTRs has also contributed significantly in the detection of total or partial chromosomal losses by the disappearance of a paternal or maternal allele on Southern blot analysis. Specifically, loss of VNTR alleles was revealed in 40 % of colorectal carcinomas from constitutionally heterozygous patients at the chromosome 17p loci, suggesting that hemi- or homo-zygosity of 17p alleles plays a role in the development of these tumors [212]. Furthermore, the detection of genetic material loss from the short arm of chromosome 17, evidenced by a VNTR allele loss, has shown that the increased p53 mRNA expression is involved in breast tumor biology [213]. Similar partial chromosomal deletions, detected with VNTR analysis, were associated with gastric carcinoma [214], neuroblastoma [215], non-Hodgkin lymphoma [216], head and neck cancer [217].

VNTRs have also played a role in the study of tumor suppressor genes, most of which were characterized by inactivation of both alleles in tumors. The use of a p53 cDNA probe and two VNTR probes on chromosome 17 has shown that the allelic loss of p53 may play an important role in the development and progression of ovarian carcinomas [218]. Furthermore, rare alleles of the H-ras VNTR, located downstream of the H-ras oncogene, have been associated with breast cancer and they have been proposed as informative markers for the breast cancer risk [219]. The use of four intragenic VNTRs as probes has shown that the loss of heterozygosity of the Rb gene is correlated with pRb protein expression and p53 alteration in human esophageal cancer [220]. Finally, a VNTR located at the intron 16 of the human retinoblastoma (RB1) gene is associated with tumor presence in retinoblastoma [221].

Many VNTRs, constituting functional promoter polymorphisms, act as genetic risk factors or pathogenic mutations for various diseases. The evaluation of a minisatellite marker expansion in the promoter of the CST6 gene has been found to be informative for the recessive myoclonus epilepsy (EPM1) onset [222], while the expansion of a VNTR upstream of the insulin gene, which regulates the insulin expression, is indicative of the polycystic ovary syndrome [223]. The alleles of another insulin gene VNTR regulatory polymorphism (26–63 repeats) predispose to type 1 diabetes in a recessive inheritance mode or are dominantly protective (140 to more than 200 repeats) [224]. Furthermore, a VNTR polymorphism of the P-selectin glycoprotein ligand-1 is a significant determinant of thrombotic predisposition in patients with antiphospholipid syndrome [225], whereas a VNTR polymorphism in the intron 2 of the interleukin-1 receptor antagonist gene is informative for the age at onset of neuropsychiatric symptoms in Wilson's disease [226].

VNTRs have also been used in prenatal diagnosis for the confirmation of parentage and the elimination of maternal contamination of chorionic villus or amniotic cell samples [227, 228]. Additionally, a VNTR identified in the human phenylalanine hydroxylase (PAH) gene has been used for the prenatal diagnosis of classical phenylketonuria [229], while prenatal diagnosis of hemophilia A can be achieved using an intragenic marker (BCL1) and an extra-genic VNTR (DXS52) [230]. Finally, VNTR genetic markers have been used in clinics to detect chimerism after bone marrow transplantation [231, 232].

Novel Methods for the Analysis of Micro- and Minisatellite DNAs

Taking into account the above, we can assume that micro- and minisatellite DNAs constitute useful genetic markers in coding and regulatory regions of the human genome. However, the role of many STRs and VNTRs remains unclear due to the technical difficulties of tandem repeat sequencing. Specifically, the next generation sequencing is probably not suitable for the analysis of genomic regions with tandem repeats because of reads with short lengths. For this reason, new methods for the analysis of micro- and minisatellite markers have been developed [233, 234].

The only limitation of these methods is their ability to genotype only repeats with short unit length and low copy numbers. Recently, a new method based on targeted enrichment for tandem repeats followed by sequencing has been developed [235], permitting a better assessment of repeat variability between individuals. The use of this method has revealed a high degree of variability in STRs and VNTRs, even between direct relatives, and the occurrence of de novo mutations. The development of such techniques will give the opportunity for new genome-wide association studies that will testify the linkage of tandem repeats with inherited or multifactorial disorders.

References

1. Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72(4):686–727
2. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44:445–477
3. Lim KG, Kwoh CK, Hsu LY, Wirawan A (2013) Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. Brief Bioinform 14(1):67–81
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al (2001) International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921
5. Szybalski W (1968) Use of cesium sulfate for equilibrium density gradient centrifugation. Methods Enzymol 12(Pt B):330–360
6. Palomeque T, Lorite P (2008) Satellite DNA, in insects: a review. Heredity 100(6):564–573
7. Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371(6494):215–220
8. Meyne J, Ratliff RL, Moyzis RK (1989) Conservation of the human telomere sequence (TTAGGG)n among vertebrates. Proc Natl Acad Sci U S A 86:7049–7056
9. Bourgain FM, Katinka MD (1991) Telomeres inhibit end to end fusion and enhance maintenance of linear DNA molecules injected into the Paramecium primaurelia macronucleus. Nucleic Acids Res 19:1541–1547
10. van Steensel B, Smorgorzewska A, de Lange T (1998) TRF2 protects human telomeres from end-to-end fusions. Cell 92:401–413
11. Ashley T (1994) Mammalian meiotic recombination: a reexamination. Hum Genet 94: 587–593
12. Gromak N, Talotti G, Proudfoot NJ, Pagani F (2008) Modulating alternative splicing by cotranscriptional cleavage of nascent intronic RNA. RNA 14(2):359–366
13. Thierry A, Bouchier C, Dujon B, Richard GF (2008) Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Res 36(18):5970–5982
14. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5(6):435–445
15. Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23:366–370
16. Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757):601–603
17. Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284(5757):604–607
18. Edelman I, Culbertson MR (1991) Exceptional codon recognition by the glutamine tRNAs in Saccharomyces cerevisiae. EMBO J 10(6):1481–1491

19. Bachtrog D, Weiss S, Zangerl B, Brem G, Schlötterer C (1999) Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. Mol Biol Evol 16(5):602–610
20. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P et al (2002) Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520–562
21. Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats: a source for the genesis of primate microsatellites. Genomics 29(1):136–144
22. Schlötterer C (1998) Genome evolution: are microsatellites really simple sequences? Curr Biol 8(4):R132–R134
23. Debrauwere H, Gendrel CG, Lechat S, Dutreix M (1997) Differences and similarities between various tandem repeat sequences: minisatellites and microsatellites. Biochimie 79(9–10):577–586
24. Pâques F, Leung WY, Haber JE (1998) Expansions and contractions in a tandem repeat induced by double-strand break repair. Mol Cell Biol 18(4):2045–2054
25. Verstrepen KJ, Jansen A, Lewitter F, Fink GR (2005) Intragenic tandem repeats generate functional variability. Nat Genet 37(9):986–990
26. Richard GF, Pâques F (2000) Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1(2):122–126
27. Gill P, Jeffreys AJ, Werrett DJ (1985) Forensic application of DNA 'fingerprints'. Nature 318(6046):577–579
28. Jeffreys AJ, Brookfield JF, Semeonoff R (1985) Positive identification of an immigration test-case using human DNA fingerprints. Nature 317(6040):818–819
29. Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. Nature 316(6023):76–79
30. Tracey M (2001) Short tandem repeat-based identification of individuals and parents. Croat Med J 42(3):233–238
31. Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28(10):1040–1050
32. Royle NJ, Clarkson RE, Wong Z, Jeffreys AJ (1988) Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. Genomics 3(4):352–360
33. Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314(6006):67–73
34. Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235(4796): 1616–1622
35. Armour JA, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR et al (1996) Minisatellite diversity supports a recent African origin for modern humans. Nat Genet 13(2):154–160
36. Bois P, Jeffreys AJ (1999) Minisatellite instability and germline mutation. Cell Mol Life Sci 55(12):1636–1648
37. Sutherland GR, Baker E, Richards RI (1998) Fragile sites still breaking. Trends Genet 14(12):501–506
38. Wahls WP, Wallace LJ, Moore PD (1990) Hypervariable minisatellite DNA, is a hotspot for homologous recombination in human cells. Cell 60(1):95–103
39. Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109(6):365–371
40. Epstein ND, Karlsson S, O'Brien S, Modi W, Moulton A, Nienhuis AW (1987) A new moderately repetitive DNA sequence family of novel organization. Nucleic Acids Res 15(5):2327–2341
41. Giacalone J, Friedes J, Francke U (1992) A novel GC-rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes. Nat Genet 1(2):137–143
42. Van Deutekom JC, Wijmenga C, van Tienhoven EA, Gruter AM, Hewitt JE, Padberg GW et al (1993) FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. Hum Mol Genet 2(12):2037–2042

43. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol 39(2):174–190
44. Gondo Y, Okada T, Matsuyama N, Saitoh Y, Yanagisawa Y, Ikeda JE (1998) Human megasatellite DNA RS447: copy-number polymorphisms and interspecies conservation. Genomics 54(1):39–49
45. Kogi M, Fukushige S, Lefevre C, Hadano S, Ikeda JE (1997) A novel tandem repeat sequence located on human chromosome 4p: isolation and characterization. Genomics 42(2):278–283
46. Saitoh Y, Miyamoto N, Okada T, Gondo Y, Showguchi-Miyata J, Hadano S et al (2000) The RS447 human megasatellite tandem repetitive sequence encodes a novel deubiquitinating enzyme with a functional promoter. Genomics 67(3):291–300
47. Okada T, Gondo Y, Goto J, Kanazawa I, Hadano S, Ikeda JE (2002) Unstable transmission of the RS447 human megasatellite tandem repetitive sequence that contains the USP17 deubiquitinating enzyme gene. Hum Genet 110(4):302–313
48. Chadwick BP (2008) DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. Genome Res 18(8):1259–1269
49. Usdin K (2008) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 18(7):1011–1019
50. Wierdl M, Greene CN, Datta A, Jinks-Robertson S, Petes TD (1996) Destabilization of simple repetitive DNA sequences by transcription in yeast. Genetics 143(2):713–721
51. Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2(8):1123–1128
52. Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62(6):1408–1415
53. Legendre M, Pochet N, Pak T, Verstrepen KJ (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 17(12):1787–1796
54. Fu YH, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S et al (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 67(6):1047–1058
55. Eichler EE, Holden JJ, Popovich BW, Reiss AL, Snow K, Thibodeau SN et al (1994) Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Nat Genet 8(1):88–94
56. Strom CM, Crossley B, Redman JB, Buller A, Quan F, Peng M et al (2007) Molecular testing for Fragile X Syndrome: lessons learned from 119,232 tests performed in a clinical laboratory. Genet Med 9(1):46–51
57. Crawford DC, Acuña JM, Sherman SL (2001) FMR1 and the fragile X syndrome: human genome epidemiology review. Genet Med 3(5):359–371
58. Hagerman PJ, Hagerman RJ (2004) Fragile X-associated tremor/ataxia syndrome (FXTAS). Ment Retard Dev Disabil Res Rev 10(1):25–30
59. Murray A (2000) Premature ovarian failure and the FMR1 gene. Semin Reprod Med 18(1):59–66
60. Sherman SL (2000) Premature ovarian failure in the fragile X syndrome. Am J Med Genet 97(3):189–194
61. Hagerman RJ (2006) Lessons from fragile X regarding neurobiology, autism, and neurodegeneration. J Dev Behav Pediatr 27(1):63–74
62. Goula AV, Merienne K (2013) Abnormal base excision repair at trinucleotide repeats associated with diseases: a tissue-selective mechanism. Genes (Basel) 4(3):375–387
63. Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11(12):2453–2465
64. Ranum LP, Day JW (2002) Dominantly inherited, non-coding microsatellite expansion disorders. Curr Opin Genet Dev 12(3):266–271

65. Karlin S, Burge C (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci U S A 93(4):1560–1565
66. Cummings CJ, Zoghbi HY (2000) Fourteen and counting: unraveling trinucleotide repeat diseases. Hum Mol Genet 9(6):909–916
67. Pearson CE, Nichol Edamura K, Cleary JD (2005) Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6(10):729–742
68. Shelbourne PF, Keller-McGandy C, Bi WL, Yoon SR, Dubeau L, Veitch NJ et al (2007) Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Hum Mol Genet 16(10):1133–1142
69. Tirosh I, Barkai N, Verstrepen KJ (2009) Promoter architecture and the evolvability of gene expression. J Biol 8(11):95
70. Brahmachari SK, Meera G, Sarkar PS, Balagurumoorthy P, Tripathi J, Raghavan S et al (1995) Simple repetitive sequences in the genome: structure and functional significance. Electrophoresis 16(9):1705–1714
71. Bacolla A, Larson JE, Collins JR, Li J, Milosavljevic A, Stenson PD et al (2008) Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res 18(10):1545–1553
72. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA et al (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One 8(2):e54710
73. Sawaya SM, Lennon D, Buschiazzo E, Gemmell N, Minin VN (2012) Measuring microsatellite conservation in mammalian evolution with a phylogenetic birth-death model. Genome Biol Evol 4(6):636–647
74. Herdewyn S, Zhao H, Moisse M, Race V, Matthijs G, Reumers J et al (2012) Whole-genome sequencing reveals a coding non-pathogenic variant tagging a non-coding pathogenic hexanucleotide repeat expansion in C9orf72 as cause of amyotrophic lateral sclerosis. Hum Mol Genet 21(11):2412–2419
75. King DG, Kashi Y (2007) Indirect selection for mutability. Heredity (Edinb) 99(2):123–124
76. Kouzine F, Levens D (2007) Supercoil-driven DNA, structures regulate genetic transactions. Front Biosci 12:4409–4423
77. Wang G, Vasquez KM (2007) Z-DNA, an active element in the genome. Front Biosci 12:4424–4438
78. Beaulieu M, Barbeau B, Rassart E (1997) Triplex-forming oligonucleotides with unexpected affinity for a nontargeted GA repeat sequence. Antisense Nucleic Acid Drug Dev 7(2):125–130
79. Rustighi A, Tessari MA, Vascotto F, Sgarra R, Giancotti V, Manfioletti G (2002) A polypyrimidine/polypurine tract within the Hmga2 minimal promoter: a common feature of many growth-related genes. Biochemistry 41(4):1229–1240
80. Han YJ, de Lanerolle P (2008) Naturally extended CT. AG repeats increase H-DNA structures and promoter activity in the smooth muscle myosin light chain kinase gene. Mol Cell Biol 28(2):863–872
81. Qin Y, Hurley LH (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. Biochimie 90(8):1149–1171
82. Du Z, Zhao Y, Li N (2009) Genome-wide colonization of gene regulatory elements by G4 DNA motifs. Nucleic Acids Res 37(20):6784–6798
83. Yadav VK, Abraham JK, Mani P, Kulshrestha R, Chowdhury S (2008) QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. Nucleic Acids Res 36(Database issue):D381–D385
84. Du Z, Zhao Y, Li N (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. Genome Res 18(2):233–241
85. Eddy J, Maizels N (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. Nucleic Acids Res 36(4):1321–1333

86. Eddy J, Vallur AC, Varma S, Liu H, Reinhold WC, Pommier Y et al (2011) G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. Nucleic Acids Res 39(12):4975–4983
87. Kumari S, Bugaut A, Huppert JL, Balasubramanian S, An RNA (2007) G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. Nat Chem Biol 3(4):218–221
88. Wieland M, Hartig JS (2007) RNA quadruplex-based modulation of gene expression. Chem Biol 14(7):757–763
89. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. Genes Dev 25(10):1010–1022
90. Halder R, Halder K, Sharma P, Garg G, Sengupta S, Chowdhury S (2010) Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. Mol Biosyst 6(12):2439–2447
91. Bacolla A, Pradhan S, Larson JE, Roberts RJ, Wells RD (2001) Recombinant human DNA (cytosine-5) methyltransferase. III. Allosteric control, reaction order, and influence of plasmid topology and triplet repeat length on methylation of the fragile X CGG.CCG sequence. J Biol Chem 276(21):18605–18613
92. Riley DE, Krieger JN (2009) UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements. Gene 429(1–2):80–86
93. Rothenburg S, Koch-Nolte F, Haag F (2001) DNA methylation and Z-DNA formation as mediators of quantitative differences in the expression of alleles. Immunol Rev 184: 286–298
94. Wells RD, Dere R, Hebert ML, Napierala M, Son LS (2005) Advances in mechanisms of genetic instability related to hereditary neurological diseases. Nucleic Acids Res 33(12): 3785–3798
95. Subirana JA, Messeguer X (2008) Structural families of genomic microsatellites. Gene 408(1–2):124–132
96. Glynn EF, Megee PC, Yu HG, Mistrot C, Unal E, Koshland DE et al (2004) Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae. PLoS Biol 2(9):E259
97. Liebich I, Bode J, Reuter I, Wingender E (2002) Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs). Nucleic Acids Res 30(15):3433–3442
98. Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P et al (1992) A second-generation linkage map of the human genome. Nature 359(6398):794–801
99. Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6(5):799–803
100. Mahtani MM, Willard HF (1993) A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci. Hum Mol Genet 2(4):431–437
101. Richards RI, Sutherland GR (1992) Dynamic mutations: a new class of mutations causing human disease. Cell 70(5):709–712
102. Kennedy GC, German MS, Rutter WJ (1995) The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription. Nat Genet 9(3):293–298
103. Turri MG, Cuin KA, Porter AC (1995) Characterisation of a novel minisatellite that provides multiple splice donor sites in an interferon-induced transcript. Nucleic Acids Res 23(11):1854–1861
104. Chaillet JR, Bader DS, Leder P (1995) Regulation of genomic imprinting by gametic and embryonic processes. Genes Dev 9(10):1177–1187
105. Neumann B, Kubicka P, Barlow DP (1995) Characteristics of imprinted genes. Nat Genet 9(1):12–13
106. Sybenga J (1999) What makes homologous chromosomes find each other in meiosis? A review and an hypothesis. Chromosoma 108(4):209–219
107. Brusco A, Saviozzi S, Cinque F, Bottaro A, DeMarchi M (1999) A recurrent breakpoint in the most common deletion of the Ig heavy chain locus (del A1-GP-G2-G4-E). J Immunol 163(8):4392–4398

108. Bennett P (2000) Demystified …microsatellites. Mol Pathol 53(4):177–183
109. Okazaki S, Tsuchida K, Maekawa H, Ishikawa H, Fujiwara H (1993) Identification of a pentanucleotide telomeric sequence, (TTAGG)n, in the silkworm Bombyxmori and in other insects. Mol Cell Biol 13(3):1424–1432
110. Wooster R, Cleton-Jansen AM, Collins N, Mangion J, Cornelis RS, Cooper CS et al (1994) Instability of short tandem repeats (microsatellites) in human cancers. Nat Genet 6(2): 152–156
111. Eshleman JR, Lang EZ, Bowerfind GK, Parsons R, Vogelstein B, Willson JK et al (1995) Increased mutation rate at the hprt locus accompanies microsatellite instability in colon cancer. Oncogene 10(1):33–37
112. Hatzistamou J, Kiaris H, Ergazaki M, Spandidos DA (1996) Loss of heterozygosity and microsatellite instability in human atherosclerotic plaques. Biochem Biophys Res Commun 225(1):186–190
113. Dubrova YE, Jeffreys AJ, Malashenko AM (1993) Mouse minisatellite mutations induced by ionizing radiation. Nat Genet 5(1):92–94
114. Jeffreys AJ, Bois P, Buard J, Collick A, Dubrova Y, Hollies CR et al (1997) Spontaneous and induced minisatellite instability. Electrophoresis 18(9):1501–1511
115. Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14(10):417–422
116. Laird CD (1990) Proposed genetic basis of Huntington's disease. Trends Genet 6(8): 242–247
117. Sabl JF, Laird CD (1992) Epigene conversion: a proposal with implications for gene mapping in humans. Am J Hum Genet 50(6):1171–1177
118. McNaught KS, Olanow CW, Halliwell B, Isacson O, Jenner P (2001) Failure of the ubiquitin-proteasome system in Parkinson's disease. Nat Rev Neurosci 2(8):589–594
119. Jakupciak JP, Wells RD (2000) Genetic instabilities of triplet repeat sequences by recombination. IUBMB Life 50(6):355–359
120. Riley DE, Krieger JN (2009) Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. Gene 429(1–2): 74–79
121. Neale MJ (2010) PRDM9 points the zinc finger at meiotic recombination hotspots. Genome Biol 11(2):104
122. Kauppi L, Jeffreys AJ, Keeney S (2004) Where the crossovers are: recombination distributions in mammals. Nat Rev Genet 5(6):413–424
123. Ségurel L, Leffler EM, Przeworski M (2011) The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. PLoS Biol 9(12):e1001211
124. Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L, May CA et al (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42(10):859–863
125. Biet E, Sun J, Dutreix M (1999) Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure. Nucleic Acids Res 27(2):596–600
126. Guo WJ, Ling J, Li P (2009) Consensus features of microsatellite distribution: microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. Genomics 93(4):323–331
127. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson BA et al (2002) A high-resolution recombination map of the human genome. Nat Genet 31(3):241–247
128. Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10(7):967–981
129. Payseur BA, Nachman MW (2000) Microsatellite variation and recombination rate in the human genome. Genetics 156(3):1285–1298

130. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321–324
131. Brandström M, Bagshaw AT, Gemmell NJ, Ellegren H (2008) The relationship between microsatellite polymorphism and recombination hot spots in the human genome. Mol Biol Evol 25(12):2579–2587
132. Varela MA, Amos W (2010) Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. Genomics 95(3): 151–159
133. Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 5:101
134. Kelkar YD, Eckert KA, Chiaromonte F, Makova KD (2011) A matter of life or death: how microsatellites emerge in and vanish from the human genome. Genome Res 21(12):2038–2048
135. Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr (2003) Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 13(6):651–658
136. Boeke JD (1997) LINEs and Alus--the polyA connection. Nat Genet 16(1):6–7
137. BatzerMA CR (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10(10):691–703
138. Dai L, Taylor MS, O'Donnell KA, Boeke JD (2012) Poly(A) binding protein C1 is essential for efficient L1 retrotransposition and affects L1 RNP formation. Mol Cell Biol 32(21):4323–4336
139. West N, Roy-Engel AM, Imataka H, Sonenberg N, Deininger P (2002) Shared protein components of SINE RNPs. J Mol Biol 321(3):423–432
140. Nadir E, Margalit H, Gallily T, Ben-Sasson SA (1996) Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proc Natl Acad Sci U S A 93(13):6470–6475
141. Clark RM, Dalgliesh GL, Endres D, Gomez M, Taylor J, Bidichandani SI (2004) Expansion of GAA triplet repeats in the human genome: unique origin of the FRDA mutation at the center of an Alu. Genomics 83(3):373–383
142. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA et al (2005) SVA elements: a hominid-specific retroposon family. J Mol Biol 354(4):994–1007
143. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73(6):1444–1451
144. McMurray CT (2010) Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet 11(11):786–799
145. La Spada AR, Taylor JP (2010) Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet 11(4):247–258
146. Dion V, Wilson JH (2009) Instability and chromatin structure of expanded trinucleotide repeats. Trends Genet 25(7):288–297
147. Pearson CE, Wang YH, Griffith JD, Sinden RR (1998) Structural analysis of slipped-strand DNA (S-DNA) formed in (CTG)n. (CAG)n repeats from the myotonic dystrophy locus. Nucleic Acids Res 26(3):816–823
148. Gacy AM, Goellner GM, Spiro C, Chen X, Gupta G, Bradbury EM et al (1998) GAA instability in Friedreich's Ataxia shares a common, DNA-directed and intraallelic mechanism with other trinucleotide diseases. Mol Cell 1(4):583–593
149. Pearson CE, Tam M, Wang YH, Montgomery SE, Dar AC, Cleary JD et al (2002) Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. Nucleic Acids Res 30(20):4534–4547
150. Pearson CE (2003) Slipping while sleeping? Trinucleotide repeat expansions in germ cells. Trends Mol Med 9(11):490–495
151. Richardson LL, Pedigo C, Ann Handel M (2000) Expression of deoxyribonucleic acid repair enzymes during spermatogenesis in mice. Biol Reprod 62(3):789–796
152. Yoon SR, Dubeau L, de Young M, Wexler NS, Arnheim N (2003) PNAS Huntington disease expansion mutations in humans can occur before meiosis is completed. Proc Natl Acad Sci U S A 100(15):8834–8838

153. Malter HE, Iber JC, Willemsen R, de Graaff E, Tarleton JC, Leisti J et al (1997) Characterization of the full fragile X syndrome mutation in fetal gametes. Nat Genet 15(2):165–169
154. Moseley ML, Schut LJ, Bird TD, Koob MD, Day JW, Ranum LP (2000) SCA8 CTG repeat: en masse contractionsin sperm and intergenerational sequence changes may play a role in reduced penetrance. Hum Mol Genet 9(14):2125–2130
155. Telenius H, Kremer B, Goldberg YP, Theilmann J, Andrew SE, Zeisler J et al (1994) Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. Nat Genet 6(4):409–414
156. Cleary JD, Tomé S, López Castel A, Panigrahi GB, Foiry L, Hagerman KA et al (2010) Tissue- and age-specific DNA replication patterns at the CTG/CAG-expanded human myotonic dystrophy type 1. Nat Struct Mol Biol 17(9):1079–1087
157. Morales F, Couto JM, Higham CF, Hogg G, Cuenca P, Braida C et al (2012) Somatic instability of the expanded CTG triplet repeat in myotonicdystrophy type 1 is a heritable quantitative trait and modifier of disease severity. Hum Mol Genet 21(16):3558–3567
158. Fischer HG, Morawski M, Brückner MK, Mittag A, Tarnok A, Arendt T (2012) Changes in neuronal DNA content variation in the human brain during aging. Aging Cell 11(4): 628–633
159. Gacy AM, Goellner G, Juranić N, Macura S, McMurray CT (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell 81(4):533–540
160. Panigrahi GB, Lau R, Montgomery SE, Leonard MR, Pearson CE (2005) Slipped (CTG)*(CAG) repeats can be correctly repaired, escape repair or undergo error-prone repair. Nat Struct Mol Biol 12(8):654–662
161. Slean MM, Reddy K, Wu B, NicholEdamura K, Kekis M, Nelissen FH et al (2013) Interconverting conformations of slipped-DNA junctions formed by trinucleotide repeats affect repair outcome. Biochemistry 52(5):773–785
162. Reddy K, Tam M, Bowater RP, Barber M, Tomlinson M, Nichol Edamura K et al (2011) Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. Nucleic Acids Res 39(5):1749–1762
163. Wheeler VC, Lebel LA, Vrbanac V, Teed A, te Riele H, MacDonald ME (2003) Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. Hum Mol Genet 12(3):273–281
164. Savouret C, Garcia-Cordier C, Megret J, te Riele H, Junien C, Gourdon G (2004) MSH2-dependent germinal CTG repeat expansions are produced continuously in spermatogonia from DM1 transgenic mice. Mol Cell Biol 24(2):629–637
165. Kovtun IV, Liu Y, Bjoras M, Klungland A, Wilson SH, McMurray CT (2007) OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. Nature 447(7143):447–452
166. Hubert L Jr, Lin Y, Dion V, Wilson JH (2011) Xpa deficiency reduces CAG trinucleotide repeat instability in neuronal tissues in a mouse model of SCA1. Hum Mol Genet 20(24):4822–4830
167. Lindahl T (2000) Suppression of spontaneous mutagenesis in human cells by DNA base excision-repair. Mutat Res 462(2–3):129–135
168. Fortini P, Dogliotti E (2007) Base damage and single-strand break repair: mechanisms and functional significance of short- and long-patch repair subpathways. DNA Repair 6(4): 398–409
169. López Castel A, Tomkinson AE, Pearson CE (2009) CTG/CAG repeat instability is modulated by the levels of human DNA ligase I and its interaction with proliferating cell nuclear antigen: a distinction between replication and slipped-DNA repair. J Biol Chem 284(39): 26631–26645
170. Tomé S, Panigrahi GB, López Castel A, Foiry L, Melton DW, Gourdon G et al (2011) Maternal germline-specific effect of DNA ligase I on CTG/CAG instability. Hum Mol Genet 20(11):2131–2143
171. van den Broek WJ, Nelen MR, van der Heijden GW, Wansink DG, Wieringa B (2006) Fen1 does not control somatic hypermutability of the (CTG)(n)*(CAG)(n) repeat in a knock-in mouse model for DM1. FEBS Lett 580(22):5208–5214

172. Møllersen L, Rowe AD, Illuzzi JL, Hildrestrand GA, Gerhold KJ, Tveterås L et al (2012) Neil1 is a genetic modifier of somatic and germline CAG trinucleotide repeat instability in R6/1 mice. Hum Mol Genet 21(22):4939–4947
173. Entezam A, Lokanga AR, Le W, Hoffman G, Usdin K (2010) Potassium bromate, a potent DNA oxidizing agent, exacerbates germline repeat expansion in a fragile X premutation mouse model. Hum Mutat 31(5):611–616
174. Goula AV, Berquist BR, Wilson DM 3rd, Wheeler VC, Trottier Y, Merienne K (2009) Stoichiometry of base excision repair proteins correlates with increased somatic CAG instability in striatum over cerebellum in Huntington's disease transgenic mice. PLoS Genet 5(12):e1000749
175. Jarem DA, Wilson NR, Delaney S (2009) Structure-dependent DNA, damage and repair in a trinucleotide repeat sequence. Biochemistry 48(28):6655–6663
176. Duval A, Hamelin R (2002) Genetic instability in human mismatch repair deficient cancers. Ann Genet 45(2):71–75
177. Shah SN, Hile SE, Eckert KA (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res 70(2):431–435
178. Coleman MG, Gough AC, Bunyan DJ, Braham D, Eccles DM, Primrose JN (2001) Minisatellite instability is found in colorectal tumours with mismatch repair deficiency. Br J Cancer 85(10):1486–1491
179. Mirkin SM (2007) Expandable DNA, repeats and human disease. Nature 447(7147):932–940
180. Swami M, Hendricks AE, Gillis T, Massood T, Mysore J, Myers RH (2009) Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. Hum Mol Genet 18(16):3039–3047
181. Drousiotou A, Stylianidou G, Anastasiadou V, Christopoulos G, Mavrikiou E, Georgiou T et al (2000) Sandhoff disease in Cyprus: population screening by biochemical and DNA analysis indicates a high frequency of carriers in the Maronite community. Hum Genet 107(1):12–17
182. Lunkes A, Trottier Y, Mandel JL (1998) Pathological mechanisms in Huntington's disease and other polyglutamine expansion diseases. Essays Biochem 33:149–163
183. Kooy RF, Willemsen R, Oostra BA (2000) Fragile X syndrome at the turn of the century. Mol Med Today 6(5):193–198
184. Mansfield ES (1993) Diagnosis of down syndrome and other aneuploidies using quantitative polymerase chain reaction and small tandem repeat polymorphisms. Hum Mol Genet 2(1):43–50
185. Zhou JL, Wei HY, Wu H, Hu YL, Liang WL (2012) Application of STR genetic marker system in the detection of hemophilia A carriers in Guangxi, China [Chinese]. Zhongguo Dang Dai Er Ke Za Zhi 14(12):951–955
186. Liu X, Wang X, Fan Q, Chu H, Fang Y, Wang H (2002) Gene diagnosis of hemophilia B by multiple STR analysis [Chinese]. Zhonghua Xue Ye Xue Za Zhi 23(3):147–150
187. Pfeifer JD, Singleton MN, Gregory MH, Lambert DL, Kymes SM (2012) Development of a decision-analytic model for the application of STR-based provenance testing of transrectal prostate biopsy specimens. Value Health 15(6):860–867
188. Jongpiputvanich S, Norapucsunton T, Mutirangura A (1996) Diagnosis and carrier detection in a Duchenne muscular dystrophy family by multiplex polymerase chain reaction and microsatellite analysis. J Med Assoc Thai 79(Suppl 1):S15–S21
189. Guzzetta V, Santoro L, Gasparo-Rippa P, Ragno M, Vita G, Caruso G et al (1995) Charcot-Marie-Tooth disease: molecular characterization of patients from central and southern Italy. Clin Genet 47(1):27–32
190. Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236(4808):1567–1570
191. Kobayashi K, Nakahori Y, Mizuno K, Miyake M, Kumagai T, Honma A et al (1998) Founder-haplotype analysis in Fukuyama-type congenital muscular dystrophy (FCMD). Hum Genet 103(3):323–327

192. Tsujikawa M, Kurahashi H, Tanaka T, Nishida K, Shimomura Y, Tano Y et al (1999) Identification of the gene responsible for gelatinous drop-like corneal dystrophy. Nat Genet 21(4):420–423

193. Riley DE, Krieger JN (2001) Short tandem repeat polymorphism linkage to the androgen receptor gene in prostate carcinoma. Cancer 92(10):2603–2608

194. Thomson JA, Pilotti V, Stevens P, Ayres KL, Debenham PG (1999) Validation of short tandem repeat analysis for the investigation of cases of disputed paternity. Forensic Sci Int 100(1–2):1–16

195. Rodig H, Roewer L, Gross A, Richter T, de Knijff P, Kayser M et al (2008) Evaluation of haplotype discrimination capacity of 35 Y-chromosomal short tandem repeat loci. Forensic Sci Int 174(2–3):182–188

196. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet 22(6):339–345

197. Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G et al (1994) Identification of the remains of the Romanov family by DNA analysis. Nat Genet 6(2):130–135

198. Clayton TM, Whitaker JP, Maguire CN (1995) Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. Forensic Sci Int 76(1):7–15

199. Kayser M, Sajantila A (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. Forensic Sci Int 118(2–3):116–121

200. Kopelman NM, Stone L, Wang C, Gefel D, Feldman MW, Hillel J et al (2009) Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. BMC Genet 10:80

201. Pope AM, Carr SM, Smith KN, Marshall HD (2011) Mitogenomic and microsatellite variation in descendants of the founder population of Newfoundland: high genetic diversity in an historically isolated population. Genome 54(2):110–119

202. Greenwood CM, Bureau A, Loredo-Osti JC, Roslin NM, Crumley MJ, Brewer CG et al (2001) Pedigree selection and tests of linkage in a Hutterite asthma pedigree. Genet Epidemiol 21(Suppl 1):S244–S251

203. Myerowitz R (2001) The search for the genetic lesion in Ashkenazi Jews with Classic Tay-Sachs disease. Adv Genet 44:137–143

204. Frankel W, Chan A, Corringham RE, Shepherd S, Rearden A, Wang-Rodriguez J (1996) Detection of chimerism and early engraftment after allogeneic peripheral blood stem cell or bone marrow transplantation by short tandem repeats. Am J Hematol 52(4):281–287

205. Molloy K, Goulden N, Lawler M, Cornish J, Oakhill A, Pamphilon D et al (1996) Patterns of hematopoietic chimerism following bone marrow transplantation for childhood acute lymphoblastic leukemia from volunteer unrelated donors. Blood 87(7):3027–3031

206. Gardiner N, Lawler M, O'Riordan J, De'Arce M, McCann SR (1997) Donor chimaerism is a strong indicator of disease free survival following bone marrow transplantation for chronic myeloid leukaemia. Leukemia 11(Suppl 3):512–515

207. Blau IW, Basara N, Serr A, Seidl C, Seifried E, Fuchs M et al (1999) A second unrelated bone marrow transplant: successful quantitative monitoring of mixed chimerism using a highly discriminative PCR-STR system. Clin Lab Haematol 21(2):133–138

208. Odelberg SJ, Plaetke R, Eldridge JR, Ballard L, O’Connell P, Nakamura Y et al (1989) Characterization of eight VNTR loci by agarose gel electrophoresis: implications for parentage testing and forensic individualization. Genomics 5:915–924

209. Raina A, Dogra TD (2002) Application of DNA fingerprinting in medicolegal practice. J Indian Med Assoc 100(12):688–694

210. Tsopanomichalou M, Sourvinos G, Arvanitis D, Michalodimitrakis M (2000) Analysis of eight polymorphic human genetic markers in a well-defined Greek population. Am J Forensic Med Pathol 21(2):172–177

211. Chakraborty R, Stivers DN, Zhong Y (1996) Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci. Mutat Res 354(1):41–48

212. Lothe RA, Nakamura Y, Woodward S, Gedde-Dahl T Jr, White R (1988) VNTR (variable number of tandem repeats) markers show loss of chromosome 17p sequences in human colorectal carcinomas. Cytogenet Cell Genet 48(3):167–169
213. Thompson AM, Steel CM, Chetty U, Hawkins RA, Miller WR, Carter DC et al (1990) p53 gene mRNA expression and chromosome 17p allele loss in breast cancer. Br J Cancer 61(1):74–78
214. Queimado L, Seruca R, Costa-Pereira A, Castedo S (1995) Identification of two distinct regions of deletion at 6q in gastric carcinoma. Genes Chromosomes Cancer 14(1):28–34
215. Ohtsu K, Hiyama E, Ichikawa T, Matsuura Y, Yokoyama T (1997) Clinical investigation of neuroblastoma with partial deletion in the short arm of chromosome 1. Clin Cancer Res 3(7):1221–1228
216. Hauptschein RS, Gamberi B, Rao PH, Frigeri F, Scotto L, Venkatraj VS et al (1998) Cloning and mapping of human chromosome 6q26-q27 deleted in B-cell non-Hodgkin lymphoma and multiple tumor types. Genomics 50(2):170–186
217. Mancini UM, Estécio MR, Góis JF, Fukuyama EE, Valentim PJ, Cury PM et al (2003) The chromosome 5q21 band minisatellite and head and neck cancer. Cancer Genet Cytogenet 147(1):87–88
218. Sakamoto T, Ogino M, Yamamoto T, Mori H, Okinaga S, Sonoda T et al (1993) Allelic losses of tumor suppressor gene on chromosome 17 in ovarian cancer [Japanese]. Nihon Sanka Fujinka Gakkai Zasshi 45(5):457–463
219. Gosse-Brun S, Sauvaigo S, Daver A, Page M, Lortholary A, Larra F et al (1999) Specific H-Ras minisatellite alleles in breast cancer susceptibility. Anticancer Res 19(6B):5191–5196
220. Xing EP, Yang GY, Wang LD, Shi ST, Yang CS (1999) Loss of heterozygosity of the Rb gene correlates with pRb protein expression and associates with p53 alteration in human esophageal cancer. Clin Cancer Res 5(5):1231–1240
221. Scharf SJ, Bowcock AM, McClure G, Klitz W, Yandell DW, Erlich HA (1992) Amplification and characterization of the retinoblastoma gene VNTR by PCR. Am J Hum Genet 50(2):371–381
222. Virtaneva K, D'Amato E, Miao J, Koskiniemi M, Norio R, Avanzini G et al (1997) Unstable minisatellite expansion causing recessively inherited myoclonus epilepsy, EPM1. Nat Genet 15(4):393–396
223. Waterworth DM, Bennett ST, Gharani N, McCarthy MI, Hague S, Batty S et al (1997) Linkage and association of insulin gene VNTR regulatory polymorphism with polycystic ovary syndrome. Lancet 349(9057):986–990
224. Bennett ST, Wilson AJ, Esposito L, Bouzekri N, Undlien DE, Cucca F et al (1997) Insulin VNTR allele-specific effect in type 1 diabetes depends on identity of untransmitted paternal allele. The IMDIAB Group. Nat Genet 17(3):350–352
225. Diz-Kucukkaya R, Inanc M, Afshar-Kharghan V, Zhang QE, López JA, Pekcelen Y (2007) P-selectin glycoprotein ligand-1 VNTR polymorphisms and risk of thrombosis in the antiphospholipid syndrome. Ann Rheum Dis 66(10):1378–1380
226. Gromadzka G, Członkowska A (2011) Influence of IL-1RN intron 2 variable number of tandem repeats (VNTR) polymorphism on the age at onset of neuropsychiatric symptoms in Wilson's disease. Int J Neurosci 121(1):8–15
227. Batanian JR, Ledbetter DH, Fenwick RG (1998) A simple VNTR-PCR method for detecting maternal cell contamination in prenatal diagnosis. Genet Test 2(4):347–350
228. Kanavakis E, Traeger-Synodinos J, Vrettou C, Maragoudaki E, Tzetis M, Kattamis C (1997) Prenatal diagnosis of the thalassaemia syndromes by rapid DNA analytical methods. Mol Hum Reprod 3(6):523–528
229. Romano V, Dianzani I, Ponzone A, Zammarchi E, Eisensmith R, Ceratto N et al (1994) Prenatal diagnosis by minisatellite analysis in Italian families with phenylketonuria. Prenat Diagn 14(10):959–962
230. Hussein IR, El-Beshlawy A, Salem A, Mosaad R, Zaghloul N, Ragab L et al (2008) The use of DNA markers for carrier detection and prenatal diagnosis of haemophilia A in Egyptian families. Haemophilia 14(5):1082–1087

231. Gatti RA, Nakamura Y, Nussmeier M, Susi E, Shan W, Grody WW (1989) Informativeness of VNTR genetic markers for detecting chimerism after bone marrow transplantation. Dis Markers 7(2):105–112
232. Kletzel M, Huang W, Olszewski M, Khan S (2013) Validation of chimerism in pediatric recipients of allogeneic hematopoietic stem cell transplantation (HSCT) a comparison between two methods: real-time PCR (qPCR) vs. variable number tandem repeats PCR (VNTR PCR). Chimerism 4(1):1–8
233. Gymrek M, Golan D, Rosset S, Erlich Y (2012) lobSTR: a short tandem repeat profiler for personal genomes. Genome Res 22:1154–1162
234. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res 41:e32
235. Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR et al (2014) Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Res 42(9):5728–5741

Chapter 8
Intron Biology, Focusing on Group II Introns, the Ancestors of Spliceosomal Introns

María Dolores Molina-Sánchez, Rafael Nisa-Martínez, Fernando M. García-Rodríguez, Francisco Martínez-Abarca, and Nicolás Toro

Characteristics of Group II Introns

Group II introns are mobile metalloribozymes that self-splice from precursor RNA to generate excised intron lariat RNA forms, which invade new DNA genomic locations by reverse splicing. These retroelements also encode a reverse transcriptase that stabilizes the RNA structure for forward and reverse splicing and finally converts the inserted intron RNA back to DNA. For these reasons, group II introns initially identified in the mitochondrial and chloroplast genomes of lower eukaryotes and plants, and subsequently found in bacteria and archaea are thought to be the ancestors of nuclear spliceosomal introns and non-long terminal repeat (non-LTR) retrotransposons [1–4]. Recently identified structural and functional similarities between group II introns and spliceosomal nuclear RNAs have suggested that group II introns may have played an important role at the very start of eukaryote evolution. It is now thought that their invasion of pre-eukaryotic genomes and their proliferation in those genomes may have driven the evolutionary separation of nucleus and cytoplasm [5].

Typically, group II introns consist of a conserved RNA structure organized into six domains. Domain V is the most conserved of these domains and is considered to be an

M.D. Molina-Sánchez, Ph.D. • R. Nisa-Martínez, Ph.D. • F.M. García-Rodríguez, Ph.D.
F. Martínez-Abarca, Ph.D. • N. Toro, Ph.D. (✉)
Grupo de Ecología Genética, Microbiología del Suelo y Sistemas Simbióticos,
Estación Experimental del Zaidín/Consejo Superior de Investigaciones Científicas,
C/Profesor Albareda, N.1, Granada, Spain 18008
e-mail: mariadolores.molina@eez.csic.es; rafael.nisa@eez.csic.es; fernando.garcia@eez.csic.es; francisco.martinez@eez.csic.es; nicolas.toro@eez.csic.es

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_8

essential part of the catalytic core (Fig. 8.1) [6, 7]. In mobile introns, a protein (the intron-encoded protein) is commonly encoded by domain IV, which contains a specific subdomain (subdomain DIVa) responsible for facilitating the canonical interaction generating the ribonucleoprotein (RNP) particles involved in the invasion of new DNA targets [8]. The IEP has two conserved domains: an N-terminal RT domain and a maturase/thumb domain (also known as the X-domain). Some IEPs also have a C-terminal DNA-binding (D) region followed by a DNA endonuclease (En) domain (Fig. 8.1).

Group II Intron Ribozyme Sequences

Group II intron ribozymes are characterized by a conserved secondary structure, which varies in size from 100 nt up to about 3000 nt [9]. The first model was established on the basis of a phylogenetic data comparison, looking for potential base pairings that had been preserved by evolution despite primary sequence divergence [6, 10]. The only sequences of group II intron ribozymes that are strongly conserved are the intron boundaries (GUGYG at the 5′ exon junction and AY at the 3′ junction), which resemble those of spliceosomal introns (GU…AG) and few nucleotides dispersed throughout the rest of the structure [11].

Group II intron ribozymes are organized into six domains, DI–DVI, radiating from a central core (Fig. 8.1). They form a structure consisting of a set of double helices resulting from Watson-Crick and Crick wobble base-pairing [12]. The six domains fold into a catalytically active tertiary structure with the assistance of a series of conserved motifs involved in long-range tertiary interactions surrounding a catalytic four-metal-ion center [13]. Some interactions involve Watson-Crick base pairs (α-α′, β-β′, γ-γ′, δ-δ′, ε-ε′, IBS1-EBS1, IBS2-EBS2 and IBS3-EBS3), whereas other are tetraloop-receptor interactions of known geometries (ζ-ζ′, η-η′ and θ-θ′), or other types of less well defined non-Watson-Crick interactions (λ-λ′, κ-κ′ and μ-μ′) [14, 15]. Two of the six domains are essential for catalysis: the largest domain (DI), and DV. Recent crystallographic studies have revealed that the shape of the RNA molecule is dictated by a set of tertiary interaction networks within domains I and III that creates the scaffold responsible for binding and activating catalytic domain V [16]. DI is essential for exon recognition in forward and reverse splicing reactions. DV contains the catalytic triad, AGC at its base, the G residue being invariant and critical for splicing. Another important catalytic motif is the AC bulge of DV. Tertiary contacts between conserved nucleotides in the linker region J2/3 and the fifth nucleotide of the intron (λ position) have been described and reported to bring together these nucleotides to form a metal ion-binding platform directly involved in catalysis [10, 17–20]. Domain VI (DVI) contains a highly conserved bulged adenosine residue that acts as the branch point for lariat formation during splicing [7]. DVI also takes part in the long-range η-η′ interaction underlying the organization of the terminal loop of DVI and an internal helix of DII, which has been reported to be important for transesterification at the 3′-splice site [9].

Group II introns can be classified into three structural subclasses, according to the recognition of their flanking exons. Subclass IIA and IIB introns, which display

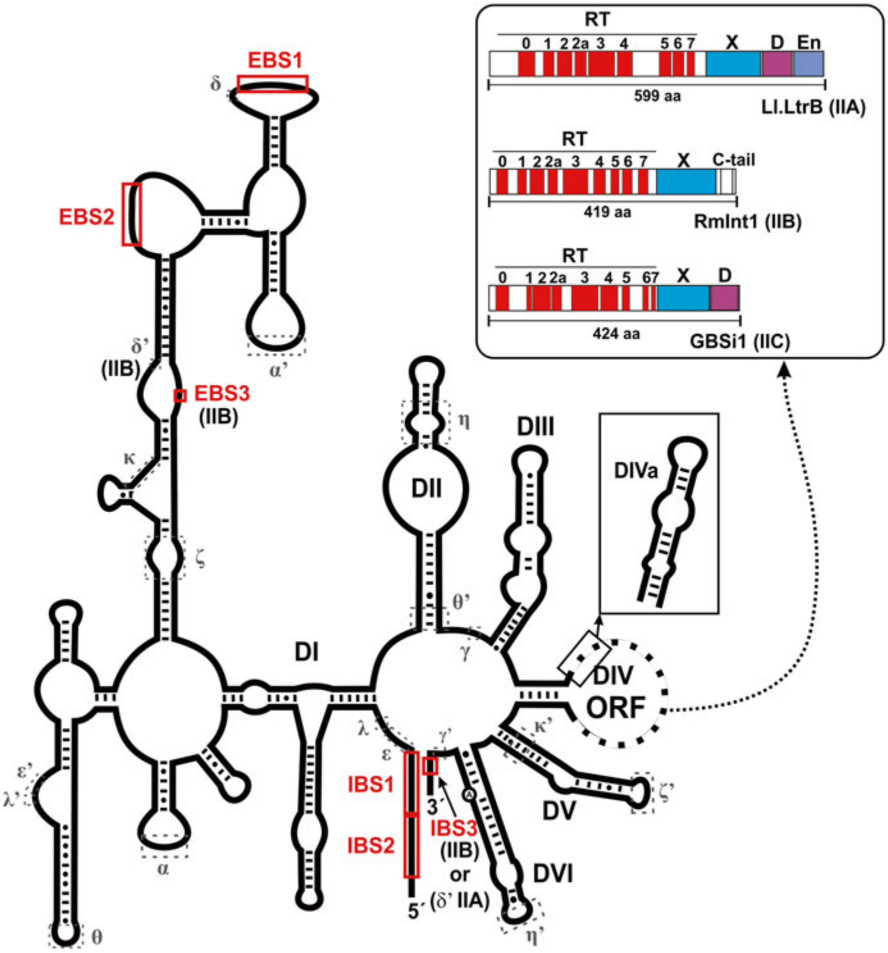


Fig. 8.1 Secondary structure of group II intron RNAs and domain structure of the intron-encoded proteins (IEP). Structure of a representative ribozyme (not to scale): domains (DI–DVI), the EBS/IBS elements, the bulged adenosine within DVI, and the sequences involved in the tertiary interactions (Greek letters) occurring during splicing, as described in the text. Major differences between group IIA and IIB introns are indicated in *brackets*. Note that group IIC introns lack the EBS2 element. The loop of DIV, which encodes the IEP, is depicted by *dashed lines*, with a box showing the location and structure of DIVa, a high-affinity binding site for the IEP. Diagrams (*drawn to scale*) of the Ll.LtrB (IIA), RmInt1 (IIB) and GBSi1 (IIC) IEPs encoded within intron DIV are boxed on the right, with the predicted reverse transcriptase (RT), maturase (X), variable DNA-binding region (D) and conserved DNA endonuclease (En) domains indicated. The C-tail denotes a C-terminal extension of 20 amino-acid residues conserved in the RmInt1 IEP and involved in maturase activity and DNA target recognition. The numbers above the RT domain identify conserved amino-acid sequence blocks characteristic of RTs

strong DNA target specificity, recognize their two exons via three base-pair interactions (IBS1/EBS1 and IBS2/EBS2 for the 5′ exon and δ-δ′ (IIA) or IBS3/EBS3 (IIB) for the 3′ exon), whereas subclass IIC introns use only two of these interactions (IBS1/EBS1 and IBS3/IBS3) and require a stem-loop structure in ssDNA that is generally derived from a transcription terminator by an as yet unknown recognition mechanism [21]. Additional subclasses have also been defined on the basis of specific structural differences: A1, A2, B1, B2 and B3 [10, 22].

Group II Intron-Encoded Proteins

The IEP component is a multifunctional protein containing a reverse transcriptase (RT) domain with subdomains conserved across other RT families (subdomains 0, 1, 2, 2a, 3, 4, 5, 6, 7) [8, 23]. Downstream from the RT domain is domain X, which functions as the thumb domain of the RT, and has a sequence conserved among group II introns but not between group II introns and other types of RTs [4, 24]. Domain X is immediately followed by a DNA-binding domain (D), defined on the basis of its function, but displaying no sequence conservation. Finally, many group II IEPs have an endonuclease domain (En) at their C-terminus that is required for the retromobility of these introns (Fig. 8.1). However, most bacterial IEPs have no En domain. Instead, they have a C-terminal portion that constitutes a distinctive, characteristic signature of this ORF class and has probably been conserved throughout evolution. This region of the protein is a group-specific functionally important protein region participating in both maturase function and intron mobility [25]. Recent studies on the consensus amino-acid sequences of the maturase and C-terminal domains have expanded our knowledge of subclasses of intron ORFs with no recognizable D/En region [26].

The classification and phylogenetic analysis of group II introns on the basis of their IEPs have resulted in the definition of several main groups: A, B, C, D, E, F, CL1 (chloroplast-like 1), CL2 (chloroplast-like 2) and ML (mitochondrion-like) [4, 26, 27], although additional types of intron ORF have recently been identified [26]. The introns of classes A, C, D, E and F and the newly identified *g1* introns encode proteins with no En domain [26]. The En domain seems to have been acquired only once in recent phylogenetic lineages of group II introns, by the common ancestor of classes B, CL and ML [26]. Intron RNA structures are generally congruent with ORF phylogeny [27].

Group II Intron Splicing

Group II introns have developed a series of splicing mechanisms to ensure their removal from the pre-mRNA and, hence, their survival in the host genome [28]. Organellar group II introns mostly interrupt essential genes, whereas prokaryotic group II introns are found in mobile genetic elements or intergenic regions [29]. Some group II introns can self-splice in vitro [30], but excision in vivo is generally assisted by protein factors that facilitate either ribozyme folding or the splicing reaction itself [31].

Excision Mechanisms of Group II Introns

Several excision mechanisms have been described for group II introns, generating different intron splicing products and ligated exons [15, 32]. These mechanisms generally co-exist within the same host [33–36], but some group II introns have been reported to display excision exclusively via a specific pathway [37–39]. Major advances have been made towards understanding the mechanism and kinetics of group II intron splicing in studies based on self-splicing assays, in which the intron precursor is incubated in non-physiological conditions (high temperatures and salt concentrations) in vitro [30].

Group II intron catalysis generally involves two sequential transesterification reactions (Fig. 8.2a). Mg^{2+} ions are involved in ensuring the correct folding of the catalytic core of the intron ribozyme and they also orchestrate the rearrangement of the single active site between the first and second splicing reactions [19, 40, 41]. The branching reaction is the most common excision pathway among group II introns [15, 42, 43] (Fig. 8.2a [1]). The 2′-OH of a bulged adenosine located in ribozyme domain VI initiates a nucleophilic attack on the phosphate bond at the intron-5′ exon junction, generating an intron intermediate in which the adenosine is covalently linked to the first intron nucleotide. In a second step, the free 3′-OH of the 5′ exon triggers a nucleophilic attack on the 3′ splice site, leading to exon ligation and intron lariat release. Alternatively, the first transesterification reaction can be initiated by a hydroxyl group from a water molecule, generating a linear excised intron and ligated exons [44–46] (Fig. 8.2a [2]). This excision pathway has classically been associated with group II introns that have lost the bulged adenosine [37–39]. Both steps in the branching reaction are reversible [47], but the first reaction of the hydrolytic pathway seems to be irreversible [33, 48]. The most recently discovered intron excision mechanism mediates the release of the intron as a true circle [35, 36, 49–51] (Fig. 8.2a [3]). The mechanism underlying this reaction remains unknown, but it has been suggested that the 3′-OH of a free 5′ exon could attack the 3′ splice site, releasing the 5′ exon-intron linear intermediate and linked exons. The hydroxyl group at the end of the intron thus reacts with the first intron residue and forms the circular intron. The maturase activity of the IEP modulates the balance between intron lariats and circles in vivo [52]. It has recently been suggested that the splicing of wheat mitochondrial group II introns under cold stress plays a regulatory role [53]. Introns with a conventional branchpoint structure display classical lariat-type splicing regardless of germination temperature. By contrast, the excision of non-conventional introns shifts from a predominantly hydrolytic pathway at room temperature to the production of circular molecules in the cold.

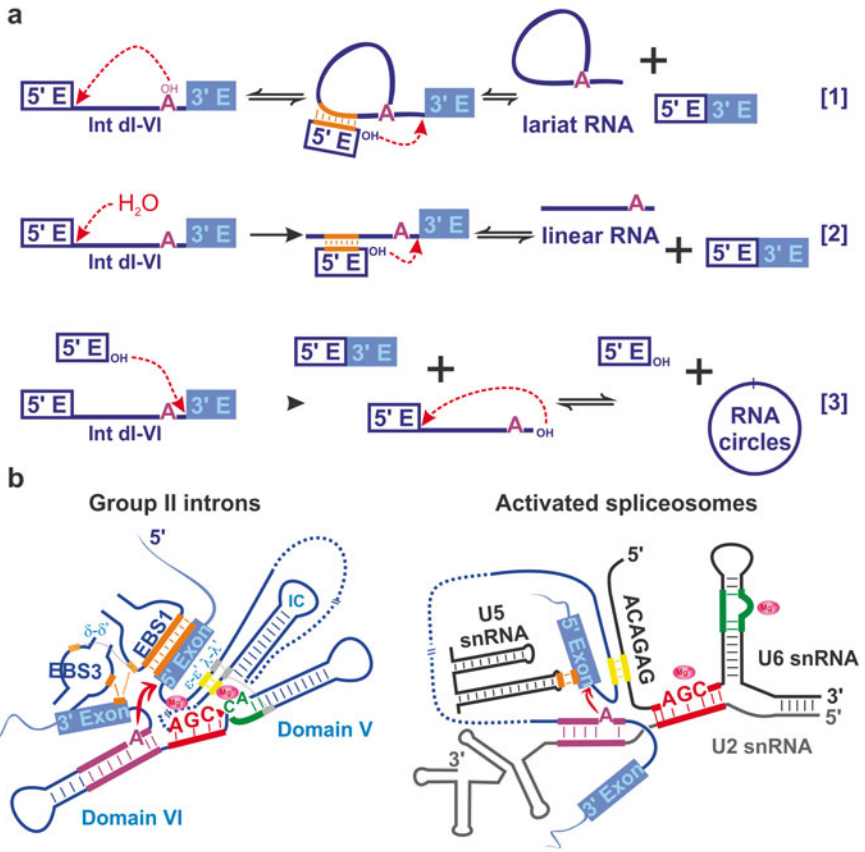


Fig. 8.2 Mechanisms of group II intron excision. (**a**) Intron lariat, linear and circular molecules may be generated, depending on the excision pathway: branching [1], hydrolysis [2] and circle formation [3], respectively. The intron RNA sequence is indicated by a *dark blue line*, and the exon sequences are represented by an *empty blue box* (5′ exon) or a *light blue box* (3′ exon). *Orange bars* correspond to the EBSs/IBSs interaction. The bulged adenosine in intron domain VI is represented as an *A*. *Dotted red lines* indicate the nucleophilic attacks occurring during each step of the reaction. (**b**) Structural similarities between the catalytic core of group II introns and activated spliceosomes. *Dark blue lines* represent the intron RNA, whereas *light blue boxes* and lines correspond to exon sequences (*dotted lines* indicating connecting regions of variable length omitted from the diagram for simplification). snRNA segments are shown in *black* or *dark gray*, and *pink spheres* represent Mg^{2+} ions. The catalytic triad is shown in *red*; the unpaired CA region in intron domain V and the equivalent segment in the U6 snRNA are shown in *green* and the bulged adenosine (A) region is shown in *violet*. The nucleophilic attack of the bulged adenosine on the 5′ exon-intron junction is indicated by a *red arrow. Orange bars* correspond to the EBSs/IBSs interaction. Other long-range contacts are identified by *yellow* and *light gray bars*

The 5′ and 3′ splice sites are recognized principally by base pairing. The EBS1-IBS1 interaction is crucial to 5′ splice site recognition during the splicing reaction. By contrast, the EBS2-IBS2 duplex is entirely dispensable for catalysis and splice site fidelity [36, 54]. Indeed, group IIC introns naturally lack the EBS2-IBS2 interaction, instead requiring the recognition of an upstream transcription terminator stem-loop [21]. Similarly, recognition of the 3′ splice site involves base pairing between the first nucleotide of the 3′ exon and the nucleotide preceding the EBS1 sequence (δ) in group IIA introns, or the EBS3 residue located in the coordination loop in group IIB and IIC introns [55, 56]. In addition to the long-range interactions underlying the catalytic conformation, tertiary contacts and structural constraints determine the efficiency and fidelity of identification, for both splice sites [11, 36, 57] (Fig. 8.2b, left panel).

Degeneration is common in organellar group II introns, for both RNA structures and ORF motifs. Thus, the intron ribozyme may be encoded in two or more pieces located at different positions in the genome, with disruptions often observed within dIV [58]. The intron and the flanking exons are transcribed separately, in a manner similar to that for *cis*-splicing introns, making the ligation of two distant exons possible. *Trans*-splicing is generally reported in higher plants, and this process requires the assistance of host-encoded protein factors [59]. A more dramatic process, intron fragmentation, has been observed in the chloroplast genome of *Euglena gracilis*, in which 155 small group II intron fragments are found [60].

Alternative splicing reactions have also been reported for bacterial and organellar group II introns [61–63]. These reactions occur at low frequency and result from misrecognition of the 3′ or 5′ splice site, inducing ORF truncations or small insertions. Alternative excision was initially thought to result in unproductive processing, but a recent study has revealed that this constitutive regulated process involving an unknown mechanism generates four functional surface-layer protein isoforms in the human pathogen *Clostridium tetani* [64].

Proteins Assisting Intron Excision

The efficiency of group II intron splicing in vivo is dependent on two groups of proteins [8]: those encoded by intron domain IV (IEPs), which are involved in *cis-splicing* reactions and found mostly in bacteria, and a group of host-encoded proteins with diverse functions, mediating the *trans*-splicing of organellar group II introns in yeasts, algae and higher plants [31, 65]. IEPs promote the splicing of group II introns in the maturase domain. They are highly specific splicing factors, playing little or no role in the excision of any intron other than the intron that encodes them [66–68]. They are usually expressed in *cis*, but they can promote the splicing of genomic copies of ORF-less introns [68, 69].

The organellar introns have diverged considerably from their bacterial ancestors, through a decrease in the number of maturase-encoding genes. A few intron-encoded ORFs are found in the mitochondria of lower eukaryotes (i.e., *Marchantia polymorpha*), but only one organelle-encoded protein has been reported to assist in the splicing of about 20 group II introns in vascular plants. This protein is called MatR in mitochondria, and MatK in chloroplasts [70, 71]. Moreover, a series of maturase-related proteins (nMat1a, nMat1b, nMat2a, and nMat2b) have been identified in the nucleus of angiosperms. After translation, these proteins are imported into mitochondria and chloroplasts to mediate the excision of group II introns [72–74]. An extensive search of the genomic sequences of recently sequenced green algae and land-plant mitochondria identified a number of new genes potentially encoding maturases, the role of which in the maturation of group II introns remains to be elucidated [75].

A plethora of nuclear-encoded proteins from various families, with diverse functions, has been reported to contribute to the correct folding and excision of organellar group II introns [9, 31, 65, 76]. One of the most numerous and well characterized groups of proteins identified is the DEAD-box proteins, which act as ATP-dependent RNA chaperones, ensuring the correct folding of the ribozyme or resolving inactive kinetic traps and then triggering productive RNA folding (the yeast factor MSS116, CYT-19 in *N. crassa*; Ded1 in *S. cerevisiae*; SrmB in *E. coli*; PMH2 in *A. thaliana*) [77–80]. Only a small number of these proteins (MSS116, CYT-19 and Ded1) can stimulate group II intron splicing in vitro in near-physiological conditions, suggesting that additional cofactors must be required to trigger splicing in vivo. Genetic and biochemical data have shown that organellar group II intron RNAs and multiple protein factors form functional high-molecular weight, spliceosome-like complexes [81, 82] (Fig. 8.2b).

Group II Intron Mobility

In addition to acting as catalytic RNAs, some group II introns are also target-specific mobile genetic elements (recently reviewed in [8, 83]). Mobile group II introns can insert site-specifically into a DNA sequence identical to the splice site (homing), at a frequency of up to 100 % [84–86], or more randomly, at low frequency, into ectopic sites (transposition) [87–89]. These mechanisms occur through an RNA intermediate and are referred to as retrohoming and retrotransposition, respectively [90, 91]. Both events occur via full reverse splicing, mediated by a ribonucleoprotein particle (RNP) formed by the association of the IEP with DIVa and the catalytic core regions of the excised intron RNA [92]. The IEP is essential for maintenance of the active intron RNA structure, to ensure that the intron can reverse splice into the DNA target site. Group II intron mobility mechanisms were first studied for the yeast mtDNA introns aI1 and aI2 [93–96], and have since been investigated in bacteria [66, 90, 97]. The main difference between yeast and bacterial mobile introns is that one or both exons may accompany the intron (co-conversion of flanking exons) in yeast, but not in bacteria [98].

The retrohoming of group II introns is highly site-specific, because a ≈20–25-bp DNA target sequence is recognized by the RNP via domain D or other regions. The IEP recognizes the upstream (positions −23 to −1) and downstream (positions +4 to +9) exon DNA sequences (Fig. 8.3a). Thirteen nucleotides of the DNA target are recognized by base pairing between the intron RNA and exon sequences, through EBS2/IBS2, EBS1/IBS1 and either δ-δ′ (IIA introns) or EBS3/IBS3 (IIB, IIC introns) interactions (Fig. 8.3b; [12, 32, 99, 100]). The essential role of each of the

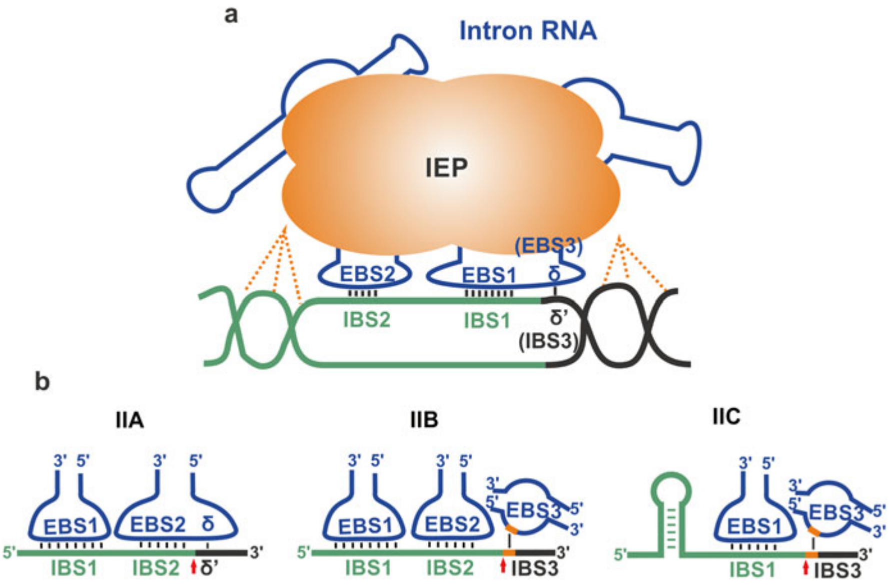


Fig. 8.3 DNA target site recognition: (**a**) RNP complexes recognize the target site on double- or single-stranded DNA primarily through EBS-IBS pairing (and by δ–δ′ interactions in subgroup IIA), whereas the IEP also binds specifically to key nucleotide residues in distal 5′ and 3′ exon regions indicated by *dotted lines* in the diagram. (**b**) Comparison between the base-pairing interactions used by group IIA, IIB and IIC introns for DNA target site recognition. *EBS* exon binding site, *IBS* intron binding site

base-pairing interactions has been demonstrated by mutating the DNA target site and observing the inhibition of reverse splicing in vitro or of intron mobility in vivo. These mutations can be rescued by compensatory mutations in the intron RNA. DNA target specificity is mostly controlled by the intron RNA, as the IEP can recognize only a few nucleotide positions. Initial recognition appears to involve interactions in the major groove between the IEP and key bases in the distal region of the 5′ exon, in the chain into which the intron subsequently reverse splices [101]. These interactions, enhanced by contact between the phosphate backbone and the IEP, involve unwinding of the DNA, allowing the reverse splicing of the intron RNA by pairing with IBS and/or δ sequences. DNA target sites have been defined experimentally for the yeast introns *coxI*-I1 [102] and *coxI*-I2 [103], the *L. lactis* Ll.LtrB intron [99], the *S. meliloti* RmInt1 intron [104], the *B. halodurans B.h.*I1 intron [21], the *E. coli E.c.*I5 intron [105], the *T. elongatus T.e.*I4h intron [106] and the *Enterobacter cloacae* group IIC *E.cl.*GOC intron [107]. Group IIA, IIB and IIC introns differ in terms of DNA target site recognition, and these differences affect design and performance in the biotechnological context (Fig. 8.3b).

Group II Intron Mobility Pathways

After reverse splicing of the RNA into the DNA target site, at least three different mechanisms may complete the mobility of yeast group II introns [96, 108]. One of these mechanisms involves a minor pathway through which a small proportion of the intron mobility events occur in natural conditions without the coconversion of exons, probably through the synthesis of a full-length intron cDNA, which is joined by DNA repair. A second RT-independent pathway (≈40 %), in which intron integration occurs by homologous recombination of both the 5′ and 3′ exon sequences, involves the repair of the nicks generated by RNPs in the DNA target, by the double-strand break reaction (DSBR) mechanism (Fig. 8.4c). Finally, the major pathway (≈60 %) entails the coconversion of the 5′ exon only, and involves cDNA synthesis by target-primed reverse transcription (TPRT) and the integration of the intron by homologous recombination (DSBR) (Fig. 8.4a, b).

The bacterial group II intron mobility pathway was first described for the *L. lactis* Ll.LtrB IIA intron, in studies involving in vivo plasmid-based genetic assays in both *L. lactis* and *E. coli* [66, 90]. Retrohoming was subsequently characterized by analyzing the biochemical characteristics of RNPs reconstituted from the purified IEP (LtrA) and in vivo excised lariat RNA [67, 109]. In vivo, the RNPs bind the DNA nonspecifically and then scan for the accurate target site by facilitated diffusion [109]. Ll.LtrB RNPs recognize a relatively long target region through three sequence motifs in DI of the RNA (EBS1, EBS2 and δ), and they base pair to complementary sequences in the DNA target site (IBS1, IBS2 and δ′) and through interactions of the IEP with nucleotides located in positions −25 to +9 of the insertion site (Fig. 8.3a) [83, 99, 100, 110, 111]. Once the target has been recognized, retrohoming occurs by TPRT (Fig. 8.4d). The RNA cleaves the sense strand of the double-stranded DNA at the exon junctions, and the intron RNA integrates into the target site. At the same time, LtrA cleaves the antisense strand at position +9, through its En activity. The 3′ end of the antisense strand is used by the RT domain of the IEP for the reverse transcription of the inserted RNA intron. The resulting cDNA is then integrated into the host DNA by homologous recombination-independent repair mechanisms [90].

Some mobile group IIB introns have an IEP with no En domain. Their IEPs have RT activity, and their RNPs can mediate reverse splicing into double- or single-stranded DNA substrates but cannot carry out site-specific second-strand cleavage; they therefore require a variant of the TPRT retrohoming pathway (Fig. 8.4e) [91, 105, 112, 113]. Most En$^-$ group II introns are inserted into the strand used as a template for synthesis of the lagging strand during replication [114]. They can therefore insert into single-stranded DNA only when the replication fork has overtaken the insertion site. The IEP thus uses the nascent lagging strand to prime reverse transcription [91]. Other mechanisms have also been suggested for the initiation of the cDNA synthesis, including random nicks in the antisense strand (*Schizosaccharomyces pombe cob*-I1 intron, [115]), or de novo initiation (RT encoded by the Mauriceville

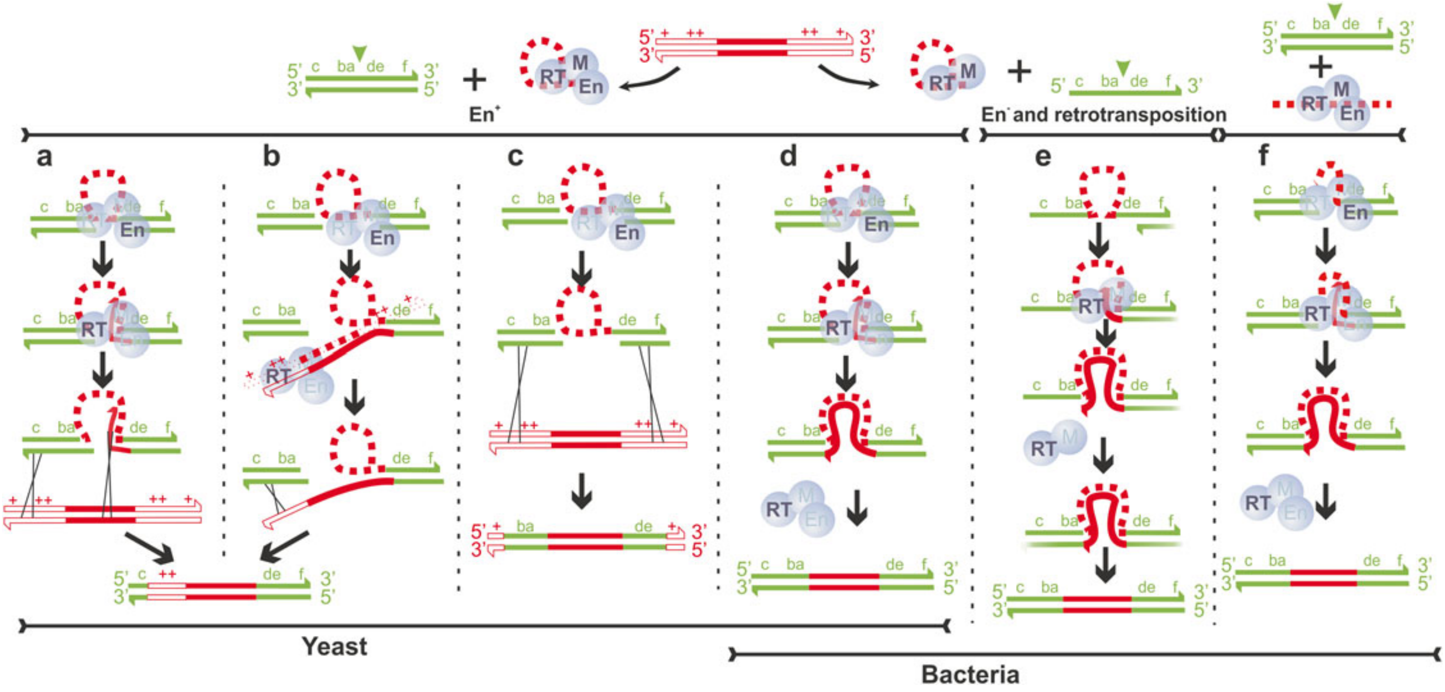


Fig. 8.4 Group II intron mobility pathways: Mechanisms *a*, *b* and *c* have been described only in yeast; all these mechanisms are dependent on homologous recombination. *a* and *b* are the major retrohoming pathways in yeast, whereas *c* is the minor pathway and is RT-independent. *d* is the major pathway in Ll.ltrB and a minor pathway in yeasts. This mechanism is independent of homologous recombination. *e* is the retrohoming pathway for introns lacking the En domain and the mechanism associated with retrotransposition. *f* is the retrohoming pathway for linear group II introns

plasmid in *Neurospora*, [116]). The best studied IIB-like intron is RmInt1 [69, 91, 112], which recognizes a DNA target site extending 20 nt into the 5′ exon and 5 nt into the 3′ exon. Target recognition occurs primarily by base-pairing between the EBS1, EBS2 and EBS3 of the intron RNA and the corresponding IBS sequences in the DNA target. The RmInt1 RT recognizes two critical nucleotide residues, possibly with the contribution of additional sequences [104, 112].

Unlike group IIA and IIB introns, group IIC introns have limited specificity due to the recognition of short IBS1 and IBS3 sequences (Fig. 8.3b). Moreover, the IBS2/EBS2 pairing seems to be replaced by the recognition of a palindromic Rho-independent transcription terminator motif or phage attachment site (*attC* sites), through an as yet unidentified mechanism [21, 37, 116–118]. Group IIC intron-encoded proteins also lack the En domain, but retain both domain Z (DNA binding) and domain X (maturase activity) [107]. Introns of this kind are found after non-identical terminators, inserted into the top or bottom strand, with a leading or inverse orientation, respectively [12, 107]. The integration of these introns resembles that of IIB introns, as it occurs through reverse splicing into single-stranded DNA at the replication fork or transcription bubble, with the nascent lagging strand preferentially used to prime reverse transcription of the intron.

It was thought that linear introns could not undergo reverse splicing [119–121], but recent studies have shown that yeast and bacterial linear group II introns reverse splice efficiently (Fig. 8.4f) [122–124]. The retrohoming of linear Ll.LtrB intron was demonstrated in eukaryotes, by the microinjection of RNPs into *Xenopus laevis* oocyte nuclei or *Drosophila melanogaster* embryos [123, 124]. The linear RNA undergoes the first reverse splicing reaction, becoming attached to the 3′ exon but not to the 5′ exon. The IEP then reverse transcribes the RNA, and the cDNA is ligated to the 5′ exon by the non-homologous end-joining (NHEJ) factor Lig 4 and the DNA repair polymerase θ (polQ). Other DNA ligases and polymerases can also perform this function, but at lower efficiency [124]. This mechanism may also mediate the retrohoming of linear RNAs, not only in eukaryotes, but also in many prokaryotes with homologous NHEJ machinery [8, 125].

Group II introns can also retrotranspose to ectopic DNA target sites, albeit at low frequency (10^{-4}–10^{-5}) [88, 114, 126–128]. The pattern of spread of Ll.LtrB within the *L. lactis* genome is consistent with intron retrotransposition into double- or single-stranded DNA through a homologous recombination-independent mechanism [114], similar to that described for the mitochondrial and bacterial RmInt1 introns [32]. In *L. lactis*, the retrotransposition of the Ll.ltrB intron is biased towards reverse splicing into transiently single-stranded DNA, with priming by the nascent lagging strand (Fig. 8.4e). By contrast, the retrotransposition of Ll.LtrB in *E. coli* is characterized by the preferential use of double-stranded DNA targets, with or without En cleavage of the opposite strand [129], indicating a role of the host cell, in addition to the intron, in pathway selection [130].

Host Factors Influencing the Retrohoming Pathway of Group II Introns

Mobile group II introns are genetic elements with specific molecular characteristics favoring their retention and spread in the genome. However, their mobility depends on the genetic background of the host, and retrohoming is dependent on the completion of cell functions [111]. The replication machinery of the cell is required in the early stages, but the host repair machinery is essential during late stages of retrohoming. The first experiments performed with the group II intron Ll.LtrB from *L. lactis* in the heterologous host *E. coli* [131] led to the formulation of a model of retrohoming involving host factors that either increased or decreased the efficiency of mobility. Thus, exonucleases (Recj, MutD, and PolI) cutting the ends of the DNA, RNases (RNase H) degrading the RNA template after cDNA synthesis, DNA and repair polymerase complexes (PolII, PolIII, PolIV and PolV) ensuring correct synthesis of the second DNA chain and DNA ligases all facilitate intron mobility. By contrast, degradative enzymes may decrease retrohoming levels. For example, RNase I and E, may eliminate the intron RNA, and exonuclease III (XthA) may degrade the newly synthesized cDNA or top strand in the upstream exon. Further studies revealed that some enzymes from the degradosome (RNase E) may affect retrohoming levels, depending on the physiological status of the cell [132]. It was subsequently shown that Ll.LtrB mobility was influenced by cell interactions and responses to cellular or environmental stresses, through global regulators [133–136]. One recent study [137] confirmed previous findings and revealed, through genetic and biochemical analyses, a possible role for replication restart proteins in the retrohoming mechanism.

Use of Group II Introns in Biotechnology

Group II introns have a number of characteristics that render them suitable for use as biotechnological tools: (1) they integrate into their DNA targets highly efficiently, in a homologous recombination-independent manner; (2) they can mobilize foreign DNA inserted within the intron; (3) minimal host functions, in the form of common cellular DNA repair mechanisms, are required for intron integration and (4) group II introns recognize the target DNA mostly through base pairing with the intron RNA. This last characteristic makes it possible to change intron specificity simply by changing the EBS/δ sequences. Currently, Ll.LtrB [138] a group IIA intron from *Lactoccocus lactis*, and the group IIB introns EcI5 [105] from *Escherichia coli* and RmInt1 from *Sinorhizobium meliloti* [139] are used as biotechnological tools. A chimeric intron based on the TeI3c ribozyme and the TeI4c IEP from *Thermosynechococcus elongatus* have also been used for gene targeting in thermophilic bacteria [140].

Introns were initially modified for the recognition of new target sites by identifying target sites matching the requirements of the IEP and then modifying the EBS/δ sequences to ensure base pairing with the new IBS/δ′ sequences. The retargeted introns are known as targetrons. Several algorithms have been developed for the retargeting of the Ll.LtrB [138], EcI5 [105] and RmInt1 [141] introns. These algorithms are based on the observed nucleotides frequencies obtained in invasion experiments using both randomized EBSs/δ-intron donor libraries and randomized IBSs/δ′-intron target site libraries. The algorithm scores a DNA sequence across a sliding window with 1 bp increments. The length of the sliding window depends on the intron: 45 bp for Ll.LtrB, 36 bp for EcI5 and 25 bp for RmInt1. A score is assigned to the potential target sites identified, with higher scores associated with a greater probability of a high invasion frequency. For each algorithm, a threshold value has been defined, above which the retargeted intron insertion frequency is high enough for the identification of intron insertion into the selected new target site in a simply assay, such as colony PCR. Once the best potential target site has been identified, the EBSs/δ sequences of the introns are modified to ensure base pairing with the new target site and are inserted into the intron donor plasmid, in which the IBSs/δ′ regions of the flanking exons are also modified to provide complementarity with the modified EBS/δ regions, for efficient RNA splicing. Intron donor plasmids also contain the IEP sequence, together with the corresponding intron from a position outside the DIV of the ribozyme (ΔORF) and downstream from the intron RNA. This conformation has been shown to be more efficient for retrohoming than the wild-type conformation with the IEP within DIV of the intron RNA [69, 105, 110]. Different promoters have been used for expression of the targetron and the associated IEP: constitutive promoters, such as the Km promoter used for the RmInt1 targetron [69], the T7 promoter recognized by the T7 RNA polymerase used in the expression of EcI5 and Ll.LtrB targetrons in *E. coli*, inducible promoters, such as the m-toluic acid-inducible promoter or tac promoter [142, 143] and endogenous promoters from the bacterial strain in which the targetron is used [144–146].

Intron integration can be detected by colony PCR or through the use of a selectable marker such as an antibiotic resistance gene. A retrotransposition-activated marker (RAM) has been developed for this purpose [145, 147]. The RAM cassette is based on a selectable marker with its own promoter inserted in reverse orientation into group II intron domain IV of the intron RNA. The marker is interrupted by the td group I intron in the forward orientation. The selectable marker is thus expressed only if retrohoming occurs. Subsequent modifications, with the selectable marker flanked by FRT sites recognized by the site-specific recombinase Flp, made it possible to remove the marker gene and led to the adaptation of the system for multiple gene disruptions.

Retargeted introns have been used in various species of the genera *Agrobacterium* [142], *Azospirillum* [148], *Bacillus* [149], *Clostridium* [145], *Ehrlichia* [150], *Escherichia* [105], *Francisella* [146], *Lactococcus* [151], *Listeria* [152], *Paenibacillus* [153], *Pasteurella* [154], *Proteus* [155], *Pseudomonas* [142], *Ralstonia* [143], *Salmonella* [111], *Shewanella* [156], *Shigella* [111], *Sinorhizobium* [139], *Sodalis* [157], *Staphylococcus* [144], *Vibrio* [158], and *Yersinia* [159]. In bacteria, targetrons are used primarily to obtain knockout mutants. In *Clostridium*, a genus in which transformation is difficult, retargeted intron derivatives of Ll.LtrB, known as ClosTron [145], have proved useful in several studies of the biology of the various species. When retargeted ΔORF introns insert into the sense strand, a conditional disruption is obtained as splicing can take place if the IEP is expressed, even *in trans*. However, intron targeting to the antisense strand leads to an unconditional mutation.

It is also possible to use targetrons to deliver foreign DNA into specific sequences [151, 160, 161]. The cargo gene is transported into the deleted region of DIV. For Ll.LtrB, fragments of less than 100 bp in length have a slight effect on intron insertion, but mobility efficiency is greatly reduced by fragments of more than 1 kb [162]. The secondary structure of the cargo sequences also affects intron mobility [156].

A method for bacterial genome editing using both targetrons and the Cre/lox system has recently been used [156]. This system has been used for insertions of 12 kb and deletions of up to 120 kb in *E. coli* and *S. aureus*, inversions in *E. coli* and *Bacillus subtilis* and one-step cut-and-paste manipulations for the translocation of 120 kb of genomic sequence to a site 1.5 Mb away.

Group II introns (Ll.LtrB) have also been used in eukaryotic cells [163]. This approach is less well developed in eukaryotes than in prokaryotes and several hurdles have yet to be overcome. The principal problem concerns the concentration of Mg^{2+}, which is below that required for the movement of Ll.LtrB in eukaryotic cells. Furthermore, the chromatinization of cellular DNA strongly inhibits intron integration. In eukaryotic cells, group II introns are microinjected into the cell nucleus as in vitro reconstituted ribonucleoproteins (RNPs). *Xenopus laevis* oocytes, and embryos of *Drosophila melanogaster* and zebrafish have been used for this purpose. RNPs have been reconstituted with both lariat and linear RNAs. In addition to the RNPs, a mixture of 500 mM Mg^{2+} and 17–20 mM each of dATP, dCTP, dGTP and dTTP is also injected into the nucleus, to optimize intron insertion. The RNPs and Mg^{2+} must be injected separately, because RNPs precipitate at this Mg^{2+} concentration. In these conditions, lariat RNPs injected into the *X. laevis* oocyte nuclei can both insert into an injected plasmid target at high frequency and stimulate DNA integration by homologous recombination, by producing target-site double-strand breaks. In *D. melanogaster* embryos, intron integration into the *yellow* gene has been achieved with introns retargeted against this gene. More knowledge is required about the behaviour of group II introns in eukaryotic cells, for the development of tools for use in eukaryotes.

Evolutionary Aspects of Bacterial Group II Introns

Group II introns display structural, functional and mechanistic similarities to eukaryotic pre-mRNA nuclear introns [164–168]. Nuclear pre-messenger RNA introns [11] and non-long terminal repeat retrotransposons may have evolved from mobile group II introns [169]. It has been suggested [2, 168, 170] that, at an early stage in the evolution of eukaryotes, the ancestral group II intron structure was split into the non-catalytic spliceosomal introns and the catalytically active RNA component of the spliceosome. This transition was accompanied by the degradation of the reverse transcriptase ORF. Maturases may have persisted in plants, during evolution, through the acquisition of a targeting signal enabling them to function within the organelles, to support the splicing of organellar group II introns [75, 171]. The evolution of eukaryotic cell organization may also have been a defensive response to the deleterious effect of group II intron proliferation in the host genome [172, 173]. Nevertheless, a recent report [174] suggests that the compartmentalization of eukaryotic cells into nucleus and cytoplasm does not prevent group II intron invasion of the host genome, although it may control proliferation of the intron, through transient or stable nucleolar sequestration. Strikingly, when the IEP loses its maturase activity, the protein becomes localized in nuclear speckles, domains of the nucleus enriched in pre-mRNA splicing factors [175], including small nuclear ribonucleoproteins (snRNPs) and serine-arginine (SR) proteins located in the interchromatin regions of the nucleoplasm. This is consistent with the hypothesis that eukaryotic spliceosomal introns may have evolved from group II introns.

Bacterial group II introns are tending to evolve towards an inactive form by fragmentation, with the loss of the 3′ terminus, including the IEP [176, 177]. The significance of fragmented introns within a particular genome remains unclear. It has recently been suggested that, as for transposable elements (TEs), the dispersal and dynamics of group II intron spread within a bacterial genome follows a selection-driven extinction model, predicting the removal of highly colonized genomes from the population by purifying selection [178]. Only 25 % of the bacterial genomes sequenced to date [8] harbor recognizable group II introns. This suggests that these introns did not act as a major force with a broad effect in the promotion of evolutionary change, but caution is required in the interpretation of these observations, because the 5′ end of fragmented intron sequences lacking the encoded ORF is unlikely to have been detected in sequenced bacterial genomes.

It is generally accepted that the "selfish" features of mobile elements underlie their acquisition and maintenance in bacterial genomes, but these elements may also be beneficial to the host. In bacteria, group II introns are thought to be tolerated to some extent because they self-splice and preferentially home to sites outside key functional genes, generally within intergenic regions or other mobile genetic elements [179]. Other studies have suggested that group II introns are beneficial to the host because they control other potentially harmful mobile genetic elements [180], and contribute to the generation of diversity and the remodeling of genomes in times of stress [135]. These features may decrease negative effects on the host organism, resulting in the maintenance of these retroelements for longer periods in bacterial populations. It also seems likely that the gradual eradication of group II introns by the host during evolution would not result in the complete elimination of intron sequences, with some intron fragments remaining and continuing to evolve in the genome. It thus remains possible that these fragments provide sequence variation on which selection can act, leading to their persistence and continuing evolution in the genomes of some bacterial lineages [181].

Acknowledgments This work was supported by research grants BIO2011-24401 and BIO2014-51953-P from the Spanish *Ministerio de Economía y Competitividad*, and CSD 2009-0006 from *Consolider-Ingenio* Programme, which includes ERDF (European Regional Development Funds).

References

1. Sharp PA (1991) Five easy pieces. Science 254:663
2. Martin W, Koonin EV (2006) Introns and the origin of nucleus-cytosol compartmentalization. Nature 440:41–45
3. Capy P, Vitalis R, Langin T, Higuet D, Bazin C (1996) Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J Mol Evol 42:359–368
4. Simon DM, Kelchner SA, Zimmerly S (2009) A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. Mol Biol Evol 26:2795–2808
5. Doolittle WF (2014) The trouble with (group II) introns. Proc Natl Acad Sci U S A 111:6536–6537
6. Michel F, Jacquier A, Dujon B (1982) Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. Biochimie 64:867–881
7. Qin PZ, Pyle AM (1998) The architectural organization and mechanistic function of group II intron structural elements. Curr Opinion Struct Biol 8:301–308
8. Lambowitz AM, Zimmerly S (2010) Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb Perspect Biol 3:a003616. doi:10.1101/cshperspect.a003616
9. Lehmann K, Schmidt U (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. Crit Rev Biochem Mol Biol 38:249–303
10. Michel F, Umesono K, Ozeki H (1989) Comparative and functional anatomy of group II catalytic introns--a review. Gene 82:5–30
11. Michel F, Ferat JL (1995) Structure and activities of group II introns. Ann Rev Biochem 64:435–461
12. Lambowitz AM, Zimmerly S (2004) Mobile group II introns. Ann Rev Genet 38:1–35
13. Marcia M, Pyle AM (2014) Principles of ion recognition in RNA: insights from the group II intron structures. RNA 20:516–527
14. Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJ et al (2008) A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase. Mol Cell 30:472–485
15. Pyle AM (2010) The tertiary structure of group II introns: implications for biological function and evolution. Crit Rev Biochem Mol Biol 45:215–232
16. Toor N, Keating KS, Fedorova O, Rajashankar K, Wang J, Pyle AM (2010) Tertiary architecture of the *Oceanobacillus iheyensis* group II intron. RNA 16:57–69
17. Pyle AM (2002) Metal ions in the structure and function of RNA. J Biol Inorg Chem 7:679–690
18. Toor N, Keating KS, Pyle AM (2009) Structural insights into RNA splicing. Curr Opinion Struct Biol 19:260–266
19. de Lencastre A, Hamill S, Pyle AM (2005) A single active-site region for a group II intron. Nat Struct Mol Biol 12:626–627
20. de Lencastre A, Pyle AM (2008) Three essential and conserved regions of the group II intron are proximal to the 5′-splice site. RNA 14:11–24
21. Robart AR, Seo W, Zimmerly S (2007) Insertion of group II intron retroelements after intrinsic transcriptional terminators. Proc Natl Acad Sci U S A 104:6620–6625
22. Toor N, Hausner G, Zimmerly S (2001) Coevolution of group II intron RNA structures with their intron-encoded reverser transcriptases. RNA 7:1142–1152

23. Zimmerly S, Hausner G, Wu X (2001) Phylogenetic relationships among group II intron ORFs. Nucleic Acids Res 29:1238–1250
24. Blocker FJ, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM (2005) Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. RNA 11:14–28
25. Molina-Sánchez MD, Martínez-Abarca F, Toro N (2010) Structural features in the C-terminal region of the *Sinorhizobium meliloti* RmInt1 group II intron-encoded protein contribute to its maturase and intron DNA-insertion function. FEBS J 277:244–254
26. Toro N, Martínez-Abarca F (2013) Comprehensive phylogenetic analysis of bacterial group II intron-encoded ORFs lacking the DNA endonuclease domain reveals new varieties. PLoS One 8:e55102
27. Toro N (2003) Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. Environ Microbiol 5:143–151
28. Pyle AM, Lambowitz AM (2006) Group II introns: ribozymes that splice RNA and invade DNA. In: Getsland RF, Cech TR, Atkins JF (eds) RNA World, 3rd edn. Cold Spring Harbor Laboratory Press, New York, pp 469–505
29. Robart AR, Zimmerly S (2005) Group II intron retroelements: function and diversity. Cytogenet Genome Res 110:589–597
30. Fedorova O, Zingler N (2007) Group II introns: structure, folding and splicing mechanism. Biol Chem 388:665–678
31. Solem A, Zingler N, Pyle AM, Li-Pook-Than J (2009) Group II introns and their protein collaborators. In: Walter NG, Woodson SA, Batey RT (eds) Non-protein coding RNAs. Springer, Berlin/Heidelberg, pp 167–182
32. Toro N, Jiménez-Zurdo JI, García-Rodríguez FM (2007) Bacterial group II introns: not just splicing. FEMS Microbiol Rev 31:342–358
33. Daniels D, Michels WJ, Pyle AM (1996) Two competing pathways for self-splicing by group II introns; a quantitative analysis of *in vitro* reaction rates and products. J Mol Biol 256:31–49
34. Costa M, Michel F, Molina-Sanchez MD, Martinez-Abarca F, Toro N (2006) An alternative intron-exon pairing scheme implied by unexpected *in vitro* activities of group II intron RmInt1 from *Sinorhizobium meliloti*. Biochimie 88:711–717
35. Molina-Sánchez MD, Martínez-Abarca F, Toro N (2006) Excision of the *Sinorhizobium meliloti* group II intron RmInt1 as circles *in vivo*. J Biol Chem 281:28737–28744
36. Nagy V, Pirakitikulr N, Zhou KI, Chillón I, Luo J, Pyle AM (2013) Predicted group II intron lineages E and F comprise catalytically active ribozymes. RNA 19:1266–1278
37. Granlund M, Michel F, Norgren M (2001) Mutually exclusive distribution of IS1548 and GBSi1, an active group II intron identified in human isolates of group B streptococci. J Bacteriol 183:2560–2569
38. Vogel J, Börner T (2002) Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. EMBO J 21:3794–3803
39. Toor N, Robart AR, Christianson J, Zimmerly S (2006) Self-splicing of a group IIC intron: 50 exon recognition and alternative 50 splicing events implicate the stem-loop motif of a transcriptional terminator. Nucleic Acids Res 34:6561–6573
40. Marcia M, Somarowthu S, Pyle AM (2013) Now on display: a gallery of group II intron structures at different stages of catalysis. Mob DNA 4:14
41. Kruschel D, Skilandat M, Sigel RKO (2014) NMR structure of the 5′ splice site in the group IIB intron Sc.ai5γ-conformational requirements for exon-intron recognition. RNA 20:295–307
42. Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL (1986) A self-splicing RNA excises an intron lariat. Cell 44:213–223
43. van der Veen R, Arnberg AC, van der Horst G, Bonen L, Tabak HF, Grivell LA (1986) Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing *in vitro*. Cell 44:225–234

44. van der Veen R, Kwakman JH, Grivell LA (1987) Mutations at the lariat acceptor site allow self-splicing of a group II intron without lariat formation. EMBO J 6:3827–3831
45. Jarrell KA, Peebles CL, Dietrich RC, Romiti SL, Perlman PS (1988) Group II intron self-splicing: alternative reaction conditions yield novel products. J Biol Chem 263: 3432–3439
46. Podar M, Chu VT, Pyle AM, Perlman PS (1998) Group II intron splicing *in vivo* by first-step hydrolysis. Nature 391:915–918
47. Chin K, Pyle AM (1995) Branch-point attack in group II introns is a highly reversible transesterification, providing a potential proofreading mechanism for 5′-splice site selection. RNA 1:391–406
48. Chu VT, Liu Q, Podar M, Perlman PS, Pyle AM (1998) More than one way to splice an RNA: branching without a bulge and splicing without branching in group II introns. RNA 4: 1186–1202
49. Murray HL, Mikheeva S, Coljee VW, Turczyk BM, Donahue WF, Bar-Shalom A, Jarrel KA (2001) Excision of group II introns as circles. Mol Cell 8:201–211
50. Li-Pook-Than J, Bonen L (2006) Multiple physical forms of excised group II intron RNAs in wheat mitochondria. Nucleic Acids Res 34:2782–2790
51. Chee GJ, Takami H (2011) Alternative splicing by participation of the group II intron ORF in extremely halotolerant and alkaliphilic *Oceanobacillus iheyensis*. Microbes Environ 26: 54–60
52. Molina-Sánchez MD, Barrientos-Durán A, Toro N (2011) Relevance of the branch point adenosine, coordination loop, and 3′ exon binding site for *in vivo* excision of the *Sinorhizobium meliloti* group II intron RmInt1. J Biol Chem 286:21154–21163
53. Dalby SJ, Bonen L (2013) Impact of low temperature on splicing of atypical group II introns in wheat mitochondria. Mitochondrion 13:647–655
54. Barrientos-Durán A, Chillón I, Martínez-Abarca F, Toro N (2011) Exon sequence requirements for excision *in vivo* of the bacterial group II intron RmInt1. BMC Mol Biol 12:24. doi:10.1186/1471-2199-12-24
55. Costa M, Michel F, Westhof E (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. EMBO J 19:5007–5018
56. Hamill S, Pyle AM (2006) The receptor for branch-site docking within a group II intron active site. Mol Cell 23:831–840
57. Su L, Qin P, Michels W, Pyle A (2001) Guiding ribozyme cleavage through motif recognition: the mechanism of cleavage site selection by a group II intron ribozyme. J Mol Biol 306:665–668
58. Bonen L (2008) *Cis*- and *trans*-splicing of group II introns in plant mitochondria. Mitochondrion 8:26–34
59. Glanz S, Kück U (2009) *Trans*-splicing of organelle introns – a detour to continuous RNAs. Bioessays 31:921–934
60. Copertino DW, Hallick RB (1993) Group II and group III introns of twintrons: potential relationship with nuclear pre-mRNA introns. Trends Biochem Sci 18:467–471
61. Jenkins K, Hong L, Hallick R (1995) Alternative splicing of the *Euglena gracilis* chloroplast *roa*A transcript. RNA 1:624–633
62. Robart AR, Montgomery NK, Smith KL, Zimmerly S (2004) Principles of 3′ splice site selection and alternative splicing for an unusual group II intron from *Bacillus anthracis*. RNA 10:854–862
63. Costa M, Michel F, Toro N (2006) Potential for alternative intron-exon pairings in group II intron RmInt1 from *Sinorhizobium meliloti* and its relatives. RNA 12:338–341
64. McNeil BA, Simon DM, Zimmerly S (2014) Alternative splicing of a group II intron in a surface layer protein gene in *Clostridium tetani*. Nucleic Acids Res 42:1959–1969
65. Brown GG, Colas des Francs-Small C, Ostersetzer-Biran O (2014) Group II intron splicing factors in plant mitochondria. Front Plant Sci 5:35. doi:10.3389/fpls.2014.00035
66. Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G et al (1997) A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemi-

cal demonstration of maturase activity and insertion of new genetic information within the intron. Genes Dev 11:2910–2924

67. Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, Lambowitz AM (1999) RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. Biochemistry 38:9069–9083
68. Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM (2004) A group II intron-encoded maturase functions preferentially *in cis* and requires both the reverse transcriptase and X domains to promote RNA splicing. J Mol Biol 340:211–231
69. Nisa-Martínez R, Jiménez-Zurdo JI, Martínez-Abarca F, Muñoz-Adelantado E, Toro N (2007) Dispersion of the RmInt1 group II intron in the *Sinorhizobium meliloti* genome upon acquisition by conjugative transfer. Nucleic Acids Res 35:214–222
70. Wahleithner JA, MacFarlane JL, Wolstenholme DR (1990) A sequence encoding a maturase-related protein in a group II intron of a plant mitochondrial *nad*1 gene. Proc Natl Acad Sci U S A 87:548–552. doi:10.1073/pnas.87.2.548
71. Zoschke R, Nakamura M, Liere K, Sugiura M, Börner T, Schmitz-Linneweber C (2010) An organellar maturase associates with multiple group II introns. Proc Natl Acad Sci U S A 107:3245–3250. doi:10.1073/pnas.0909400107
72. Nakagawa N, Sakurai N (2006) A mutation in *At-nMat1a*, which encodes a nuclear gene having high similarity to group II intron maturase, causes impaired splicing of mitochondrial *nad*4 transcript and altered carbon metabolism in *Arabidopsis thaliana*. Plant Cell Physiol 47:772–783. doi:10.1093/pcp/pcj051
73. Keren I, Bezawork-Geleta A, Kolton M, Maayan I, Belausov E, Levy M et al (2009) AtnMat2, a nuclear-encoded maturase required for splicing of group-II introns in *Arabidopsis* mitochondria. RNA 15:2299–2311. doi:10.1261/rna.1776409
74. Cohen S, Zmudjak M, Colas des Francs-Small C, Malik S, Shaya F, Keren I et al (2014) nMAT4, a maturase factor required for *nad*1 pre-mRNA processing and maturation, is essential for holocomplex I biogenesis in *Arabidopsis* mitochondria. Plant J 78:253–268
75. Guo W, Mower J (2013) Evolution of plant mitochondrial intron-encoded maturases: frequent lineage-specific loss and recurrent intracellular transfer to the nucleus. J Mol Evol 77:43–54. doi:10.1007/s00239-013-9579-7
76. de Longevialle AF, Small ID, Lurin C (2010) Nuclearly encoded splicing factors implicated in RNA splicing in higher plant organelles. Mol Plant 3:691–705. doi:10.1093/mp/ssq025
77. Huang HR, Rowe CE, Mohr S, Jiang Y, Lambowitz AM, Perlman PS (2005) The splicing of yeast mitochondrial group I and group II introns requires a DEAD-box protein with RNA chaperone function. Proc Natl Acad Sci U S A 102:163–168
78. Halls C, Mohr S, del Campo M, Yang Q, Jankowsky E, Lambowitz AM (2007) Involvement of DEAD-box proteins in group I and group II intron splicing: biochemical characterization of Mss116p, ATP hydrolysis-dependent and -independent mechanisms, and general RNA chaperone activity. J Mol Biol 365:835–855
79. del Campo M, Mohr S, Jiang Y, Jia H, Jankowsky E, Lambowitz AM (2009) Unwinding by local strand separation is critical for the function of DEAD-box proteins as RNA chaperones. J Mol Biol 389:674–693
80. Köhler D, Schmidt-Gattung S, Binder S (2010) The DEAD-box protein PMH2 is required for efficient group II intron splicing in mitochondria of *Arabidopsis thaliana*. Plant Mol Biol 72:459–467. doi:10.1007/s11103-009-9584-9
81. Jacobs J, Glanz S, Bunse-Graβmann A, Kruse O, Kück U (2010) RNA *trans*-splicing: identification of components of a putative chloroplast spliceosome. Eur J Cell Biol 89:932–939
82. Jacobs J, Marx C, Kock V, Reifschneider O, Fränzel B, Krisp C et al (2013) Identification of a chloroplast ribonucleoprotein complex containing *trans*-splicing factors, intron RNA, and novel components. Mol Cell Proteomics 12:1912–1925. doi:10.1074/mcp.M112.026583
83. Enyeart PJ, Mohr G, Ellington AD, Lambowitz AM (2014) Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and noncoding RNA analysis. Mob DNA 5:2. doi:10.1186/1759-8753-5-2

84. Skelly PJ, Hardy CM, Clark-Walker GD (1991) A mobile group II intron of a naturally occurring rearranged mitochondrial genome in *Kluyveromyces lactis*. Curr Genet 20:115–120
85. Lazowska J, Meunier B, Macadre C (1994) Homing of a group II intron in yeast mitochondrial DNA is accompanied by unidirectional co-conversion of upstream-located markers. EMBO J 13:4963–4972
86. Moran JV, Zimmerly S, Eskes R, Kennell JC, Lambowitz AM, Butow RA, Perlman PS (1995) Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. Mol Cell Biol 15:2828–2838
87. Mueller MW, Allmaier M, Eskes R, Schweyen RJ (1993) Transposition of group II intron aI1 in yeast and invasion of mitochondrial genes at new locations. Nature 366:174–176
88. Cousineau B, Lawrence S, Smith D, Belfort M (2000) Retrotransposition of a bacterial group II intron. Nature 404:1018–1021 (Erratum: Nature 2001;414:84)
89. Muñoz E, Villadas PJ, Toro N (2001) Ectopic transposition of a group II intron in natural bacterial populations. Mol Microbiol 41:645–652
90. Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D et al (1998) Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. Cell 94:451–462
91. Martínez-Abarca F, Barrientos-Durán A, Fernández-López M, Toro N (2004) The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. Nucleic Acids Res 32:2880–2888
92. Lambowitz AM, Mohr G, Zimmerly S (2005) Group II intron homing endonucleases: ribonucleoprotein complexes with programmable target specificity. In: Belfort M, Stoddard BL, Wood DW, Derbyshire V (eds) Homing endonucleases and inteins. Springer, Heidelberg, pp 121–145
93. Zimmerly S, Guo H, Perlman PS, Lambowitz AM (1995) Group II intron mobility occurs by target DNA-primed reverse transcription. Cell 82:545–554
94. Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM (1995) A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. Cell 83:529–538
95. Yang J, Zimmerly S, Perlman PS, Lambowitz AM (1996) Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. Nature 381:332–335
96. Eskes R, Yang J, Lambowitz AM, Perlman PS (1997) Mobility of yeast mitochondrial group II introns: engineering a new site specificity and retrohoming via full reverse splicing. Cell 88:865–874
97. Mills DA, Manias DA, McKay LL, Dunny GM (1997) Homing of a group II intron from *Lactococcus lactis* subsp. lactis ML3. J Bacteriol 179:6107–6111
98. Belfort M, Derbyshire V, Parker MM, Cousineau B, Lambowitz AM (2002) Mobile introns: pathways and proteins. In: Craig NL, Gragie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, DC, pp 761–781
99. Mohr G, Smith D, Belfort M, Lambowitz AM (2000) Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. Genes Dev 14:559–573
100. Singh NN, Lambowitz AM (2001) Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. J Mol Biol 309:361–386
101. Singh RN, Saldanha RJ, D'Souza LM, Lambowitz AM (2002) Binding of a group II intron-encoded reverse transcriptase/maturase to its high-affinity intron RNA binding site involves sequence-specific recognition and autoregulates translation. J Mol Biol 318:287–303
102. Yang J, Mohr G, Perlman PS, Lambowitz AM (1998) Group II intron mobility in yeast mitochondria: target DNA-primed reverse transcription activity of aI1 and reverse splicing into DNA transposition sites *in vitro*. J Mol Biol 282:505–523
103. Guo H, Zimmerly S, Perlman PS, Lambowitz AM (1997) Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. EMBO J 16:6835–6848

104. Jiménez-Zurdo JI, García-Rodríguez FM, Barrientos-Durán A, Toro N (2003) DNA target site requirements for homing *in vivo* of a bacterial group II intron encoding a protein lacking the DNA endonuclease domain. J Mol Biol 326:413–423
105. Zhuang F, Karberg M, Perutka J, Lambowitz AM (2009) EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. RNA 15:432–449
106. Mohr G, Ghanem E, Lambowitz AM (2010) Mechanisms used for genomic proliferation by thermophilic group II introns. PLoS Biol 8:e1000391. doi:10.1371/journal.pbio.1000391
107. Rodríguez-Martínez JM, Nordmann P, Poirel L (2012) Group IIC intron with an unusual target of integration in *Enterobacter cloacae*. J Bacteriol 194:150–160. doi:10.1128/JB.05786-11
108. Eskes R, Liu L, Ma H, Chao MY, Dickson L, Lambowitz AM, Perlman PS (2000) Multiple homing pathways used by yeast mitochondrial group II introns. Mol Cell Biol 20:8432–8446
109. Aizawa Y, Xiang Q, Lambowitz AM, Pyle AM (2003) The pathway of DNA recognition and RNA integration by a group II intron retrotransposon. Mol Cell 11:795–805
110. Guo H, Karberg M, Long M, Jones JP 3rd, Sullenger B, Lambowitz AM (2000) Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. Science 289:452–457
111. Karberg M, Guo H, Zhong J, Coon R, Perutka J, Lambowitz AM (2001) Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nat Biotechnol 19:1162–1167
112. Muñoz-Adelantado E, San Filippo J, Martínez-Abarca F, García-Rodríguez FM, Lambowitz AM, Toro N (2003) Mobility of the *Sinorhizobium meliloti* group II intron RmInt1 occurs by reverse splicing into DNA, but requires an unknown reverse transcriptase priming mechanism. J Mol Biol 327:931–943
113. Zhong J, Lambowitz AM (2003) Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. EMBO J 22:4555–4565
114. Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M (2002) Multiple pathways for the Ll.LtrB group II intron include reverse splicing into DNA targets. Mol Microbiol 46:1259–1271
115. Schäfer B, Gan L, Perlman PS (2003) Reverse transcriptase an reverse splicing activities encoded by the mobile group II intron *COB*I1 of fission yeast mitochondrial DNA. J Mol Biol 329:191–206
116. Centron D, Roy PH (2002) Presence of a group II intron in a multiresistant *Serratia marcescens* strain that harbors three integrons and a novel gene fusion. Antimicrob Agents Chemother 46:1402–1409
117. Dai L, Zimmerly S (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. Nucleic Acids Res 30:1091–1102
118. Léon G, Roy PH (2009) Group IIC intron mobility into *attC* sites involves a bulged DNA stem-loop motif. RNA 15:1543–1553
119. Dème E, Nolte A, Jacquier A (1999) Unexpected metal ion requirements specific for catalysis of the branching reaction in a group II intron. Biochemistry 38:3157–3167
120. Mohr S, Matsuura M, Perlman PS, Lambowitz AM (2006) A DEAD-box protein alone promotes group II intron splicing and reverse splicing by acting as an RNA chaperone. Proc Natl Acad Sci U S A 103:3569–3574
121. Gordon PM, Fong R, Piccirilli JA (2007) A second divalent metal ion in the group II intron reaction center. Chem Biol 14:607–612
122. Roitzsch M, Pyle AM (2009) The linear form of a group II intron catalyzes efficient autocatalytic reverse splicing, establishing a potential for mobility. RNA 15:473–482
123. Zhuang F, Mastroianni M, White TB, Lambowitz AM (2009) Linear group II intron RNAs can retrohome in eukaryotes and may use nonhomologous end-joining for cDNA ligation. Proc Natl Acad Sci U S A 106:18189–18194

124. White TB, Lambowitz AM (2012) The retrohoming of linear group II intron RNAs in *Drosophila melanogaster* occurs by both DNA ligase 4-dependent and -independent mechanisms. PLoS Genet 8:e1002534. doi:10.1371/journal.pgen.1002534
125. Bowater R, Doherty AJ (2006) Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. PLoS Genet 2:e8. doi:10.1371/journal.pgen.0020008
126. Martínez-Abarca F, Toro N (2000) RecA-independent ectopic transposition *in vivo* of a bacterial group II intron. Nucleic Acids Res 28:4397–4402
127. Dickson L, Huang HR, Liu L, Matsuura M, Lambowitz AM, Perlman PS (2001) Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. Proc Natl Acad Sci U S A 98:13207–13212
128. Ichiyanagi K, Beauregard A, Belfort M (2003) A bacterial group II intron favors retrotransposition into plasmid targets. Proc Natl Acad Sci U S A 100:15742–15747
129. Coros CJ, Landthaler M, Piazza CL, Beauregard A, Esposito D, Perutka J (2005) Retrotransposition strategies of the *Lactococcus lactis* Ll.LtrB group II intron are dictated by host identity and cellular environment. Mol Microbiol 56:509–524
130. Beauregard A, Chalamcharla VR, Piazza CL, Belfort M, Coros CJ (2006) Bipolar localization of the group II intron Ll.LtrB is maintained in *Escherichia coli* deficient in nucleoid condensation, chromosome partitioning and DNA replication. Mol Microbiol 62:709–722
131. Smith D, Zhong J, Matsuura M, Lambowitz AM, Belfort M (2005) Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. Genes Dev 19:2477–2487
132. Coros CJ, Piazza CL, Chalamcharla VR, Belfort M (2008) A mutant screen reveals RNase E as a silencer of group II intron retromobility in *Escherichia coli*. RNA 14:2634–2644
133. Beauregard A, Curcio MJ, Belfort M (2008) The take and give between retrotransposable elements and their hosts. Ann Rev Genet 42:587–617
134. Zhao J, Niu W, Marcotte E, Lambowitz A (2008) Group II intron protein localization and insertion sites are affected by polyphosphate. PLoS Biol 6:e150. doi:10.1371/journal.pbio.0060150
135. Coros CJ, Piazza CL, Chalamcharla VR, Smith D, Belfort M (2009) Global regulators orchestrate group II intron retromobility. Mol Cell 34:250–256
136. Edgell DR, Chalamcharla VR, Belfort M (2011) Learning to live together: mutualism between self-splicing introns and host genomes. BMC Biol 9:22. doi:10.1186/1741-7007-9-22
137. Yao J, Truong DM, Lambowitz AM (2013) Genetic and biochemical assays reveal a key role for replication restart proteins in group II intron retrohoming. PLoS Genet 9:e1003469. doi:10.1371/journal.pgen.1003469
138. Perutka J, Wang W, Goerlitz D, Lambowitz AM (2004) Use of computer-designed group II introns to disrupt *Escherichia coli* DExH/D-box protein and DNA helicase genes. J Mol Biol 336:421–439
139. García-Rodríguez FM, Barrientos-Durán A, Díaz-Prado V, Fernández-López M, Toro N (2011) Use of RmInt1, a group IIB intron lacking the intron-encoded protein endonuclease domain, in gene targeting. Appl Environ Microbiol 77:854–861. doi:10.1128/AEM.02319-10
140. Mohr G, Hong W, Zhang J, Cui GZ, Yang Y, Cui Q et al (2013) A targetron system for gene targeting in thermophiles and its application in *Clostridium thermocellum*. PLoS One 8:e69032. doi:10.1371/journal.pone.0069032
141. García-Rodríguez FM, Hernández-Gutiérrez T, Díaz-Prado V, Toro N (2014) Use of the computer-retargeted group II intron RmInt1 of *Sinorhizobium meliloti* for gene targeting. RNA Biol 11:391–401
142. Yao J, Lambowitz AM (2007) Gene targeting in gram-negative bacteria by use of a mobile group II intron (“targetron”) expressed from a broad-host-range vector. Appl Environ Microbiol 73:2735–2743
143. Park JM, Jang YS, Kim TY, Lee SY (2010) Development of a gene knockout system for *Ralstonia eutropha* H16 based on the broad-host-range vector expressing a mobile group II intron. FEMS Microbiol Lett 309:193–200

144. Yao J, Zhong J, Fang Y, Geisinger E, Novick RP, Lambowitz AM (2006) Use of targetrons to disrupt essential and nonessential genes in *Staphylococcus aureus* reveals temperature sensitivity of Ll.LtrB group II intron splicing. RNA 12:1271–1281
145. Heap JT, Pennington OJ, Cartman ST, Carter GP, Minton NP (2007) The ClosTron: a universal gene knock-out system for the genus *Clostridium*. J Microbiol Methods 70:452–464
146. Rodriguez SA, Yu JJ, Davis G, Arulanandam BP, Klose KE (2008) Targeted inactivation of *Francisella tularensis* genes by group II introns. Appl Environ Microbiol 74:2619–2626
147. Zhong J, Karberg M, Lambowitz AM (2003) Targeted and random bacterial gene disruption using a group II intron (tagetron) vector containing a retrotransposition-activated selectable marker. Nucleic Acids Res 31:1656–1664
148. Malhotra M, Srivastava S (2008) An *ipdC* gene knock-out of *Azospirillum brasilense* strain SM and its implications on indole-3-acetic acid biosynthesis and plant growth promotion. Antonie Van Leeuwenhoek 93:425–433
149. Akhtar P, Hkan SA (2012) Two independent replicons can support replication of the anthrax toxin-encoding plasmid pXO1 of *Bacillus anthracis*. Plasmid 67:111–117
150. Cheng C, Nair AD, Indukuri W, Gong S, Felsheim RF, Jaworski D et al (2013) Targeted and random mutagenesis of *Ehrlichia chaffeensis* for the identification of genes required for *in vivo* infection. PLoS Pathog 9:e1003171. doi:10.1371/journal.ppat.1003171
151. Frazier CL, San Filippo J, Lambowitz AM, Mills DA (2003) Genetic manipulation of *Lactococcus lactis* by using targeted group II introns: generation of stable insertions without selection. Appl Environ Microbiol 69:1121–1128
152. Alonzo F 3rd, Port GC, Cao M, Freitag NE (2009) The postranslocation chaperone PrsA2 contributes to multiple facets of *Listeria monocytogenes* pathogenesis. Infect Immun 77:3077–3085
153. Zarschler K, Janesch B, Zayni S, Schäffer C, Messner P (2009) Construction of a gene knockout system for application in *Paenibacillus alvei* CCM 2051 T, exemplified by the S-layer glycan biosynthesis initiation enzyme WsfP. Appl Environ Microbiol 75:3077–3085
154. Steen JA, Steen JA, Harrison P, Seemann T, Wilkie I, Harper M, Adler B, Boyce JD (2010) Fis is essential for capsule production in *Pasteurella multocida* and regulates expression of other important virulence factors. PLoS Pathog 6:e1000750. doi:10.1371/journal.ppat.1000750
155. Pearson MM, Mobley HL (2007) The type III secretion system of *Proteus mirabilis* HI4320 does not contribute to virulence in the mouse model of ascending urinary tract infection. J Med Microbiol 56:1277–1283
156. Enyeart PJ, Chirieleison SM, Dao MN, Perutka J, Quandt EM, Yao J et al (2013) Generalized bacterial genome editing using mobile group II introns and Cre-lox. Mol Syst Biol 9:685. doi:10.1038/msb.2013.41
157. Smith CL, Weiss BL, Aksoy S, Runyen-Janecky LJ (2013) Characterization of the achromobactin iron acquisition operon in *Sodalis glossinidius*. Appl Environ Microbiol 79:2872–2881
158. Kumar S, Smith KP, Floyd JL, Varela MF (2011) Cloning and molecular analysis of a mannitol operon of phosphoenolpyruvate-dependent phosphotransferase (PTS) type from *Vibrio cholerae* O395. Arch Microbiol 193:201–208
159. Palonen E, Lindstrom M, Karttunen R, Somervuo P, Korkeala H (2011) Expression of signal transduction system encoding genes of *Yersinia pseudotuberculosis* IP32953 at 28°C and 3°C. PLoS One 6:e25063. doi:10.1371/journal.pone.0025063
160. Jones JP III, Kierlin MN, Coon RB, Perutka J, Lambowitz AM, Sullenger BA (2005) Retargeting mobile group II introns to repair mutant genes. Mol Ther 11:687–694
161. Rawsthorne H, Turner KN, Mills DA (2006) Multicopy integration of heterologous genes, using the lactococcal group II intron targeted to bacterial insertion sequences. Appl Environ Microbiol 72:6088–6093
162. Plante I, Cousineau B (2006) Restriction for gene insertion within the *Lactococcus lactis* Ll.LtrB group II intron. RNA 12:1980–1992
163. Mastroianni M, Watanabe K, Whit TB, Zhuang F, Vernon J, Matsuura M, Wallingford J, Lambowitz AM (2008) Group II intron-based gene targeting reactions in eukaryotes. PLoS One 3:e3121. doi:10.1371/journal.pone.0003121

164. Grabowski PJ, Seiler SA, Sharp PA (1985) A multicomponent complex is involved in the splicing of messenger RNA precursors. Cell 42:345–353
165. Cech TK (1986) The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell 44:207–210
166. Patel AA, Steitz JA (2003) Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 4:960–970
167. Toor N, Keating KS, Taylor SD, Pyle AM (2008) Crystal structure of a self-spliced group II intron. Science 320:77–82
168. Keating KS, Toor N, Perlman PS, Pyle AM (2010) A structural analysis of the group II intron active site and implications for the spliceosome. RNA 16:1–9
169. Eickbush TH (1994) Origins and evolutionary relationships of retroelements. In: Morse SS (ed) The evolutionary biology of viruses. Raven, New York, pp 121–157
170. Rogozin IB, Carmel L, Csuros M, Koonin EV (2012) Origin and evolution of spliceosomal introns. Biol Direct 7:11. doi:10.1186/1745-6150-7-11
171. Mohr G, Lambowitz AM (2003) Putative proteins related to group II intron reverse transcriptase/maturases are encoded by nuclear genes in higher plants. Nucleic Acids Res 31:647–652
172. Koonin EV (2006) The origin of introns and the role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct 1:22. doi:10.1186/1745-6150-1-22
173. Koonin EV (2011) The origins of eukaryotes: endosymbiosis, the strange story of introns, and the ultimate importance of unique events in evolution. In: Pearson Education Inc. (ed) The logic of chance: the nature and origin of biological evolution, 1st edn. FT Press Science, Upper Saddle River, pp 171–224
174. Nisa-Martínez R, Laporte P, Jiménez-Zurdo JI, Frugier F, Crespi M, Toro N (2013) Localization of a bacterial group II intron-encoded protein in eukaryotic nuclear splicing-related cell compartments. PLoS One 8(12):e84056. doi:10.1371/journal.pone.0084056, eCollection 2013
175. Spector DL, Lamond AI (2011) Nuclear speckles. Cold Spring Harb Perspect Biol 3:a000646. doi:10.1101/cshperspect.a000646
176. Dai LX, Zimmerly S (2002) The dispersal of five group II introns among natural populations of *Escherichia coli*. RNA 8:1294–1307
177. Fernandez-Lopez M, Munoz-Adelantado E, Gillis M, Willems A, Toro N (2005) Dispersal and evolution of the *Sinorhizobium meliloti* group II RmInt1 intron in bacteria that interact with plants. Mol Biol Evol 22:1518–1528
178. Leclercq S, Cordaux R (2012) Selection-driven extinction dynamics for group II introns in Enterobacteriales. PLoS One 7:e52268
179. Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, Dai L et al (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. RNA 14:1704–1713
180. Chillón I, Martínez-Abarca F, Toro N (2010) Splicing of the *Sinorhizobium meliloti* RmInt1 group II intron provides evidence of retroelement behavior. Nucleic Acids Res 39:1095–1104
181. Toro N, Martínez-Rodríguez L, Martínez-Abarca F (2014) Insights into the history of a bacterial group II intron remnant from the genomes of the nitrogen-fixing symbionts *Sinorhizobium meliloti* and *Sinorhizobium medicae*. Heredity. doi:10.1038/hdy.2014.32

Chapter 9
Centromeres in Health, Disease and Evolution

Thian Thian Beh and Paul Kalitsis

Abbreviations

ANIFAD	Associated network instability, and facial anomalies division
ALP-WDLPS	Atypical lipomas and well-differentiated liposarcomas
AML	Acute myeloid leukemia
ATLL	Adult T-cell leukemia/lymphoma
CCAN	Constitutive centromere-associated network
CENP	Centromere protein
CGI	CpG island
CIN	Chromosomal instability
CLL	Chronic lymphocytic leukemia
CPC	Chromosome passenger complex
FISH	Fluorescence in situ hybridization
HOR	Higher order repeat
ICF	Immunodeficiency, centromeric instability, and facial anomalies
MVA	Mosaic variegated aneuploidy
PCD	Premature centromere division
PCS	Premature chromatid separation
ROB	Robertsonian translocation
SAC	Spindle assembly checkpoint
TE	Transposable element

T.T. Beh, B.Sc. • P. Kalitsis, Ph.D. (✉)
Murdoch Childrens Research Institute, Royal Children's Hospital,
Parkville, VIC 3052, Australia

Department of Paediatrics, University of Melbourne, Melbourne, VIC 3052, Australia
e-mail: thianthian.beh@mcri.edu.au; paul.kalitsis@mcri.edu.au

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_9

Introduction

The term centromere (*kentron*, center; *meros*, part), initially coined by Waldeyer in 1903 for the neck of sperm, was reinterpreted by Darlington in 1936 as the centric constriction on metaphase chromosomes to which spindle fibers attach during cell division [1]. Centromeres were cytologically distinguished by their constricted morphological appearance and with C-banding which is a Giemsa staining procedure that preferentially stains the heterochromatin regions [2]. Now, fluorescence in situ hybridization (FISH) probes against centromeric DNA of specific chromosome and antibodies against centromere proteins are commonly used for the localization of centromeric regions [3–5].

The human centromere is a region on the chromosome consisting of an underlying alpha satellite repetitive DNA sequence that winds around nucleosomes containing centromere protein (CENP)-A, a histone H3 variant. Hence, CENP-A is the epigenetic mark of a centromere and is one of the 17 proteins forming the constitutive centromere-associated network (CCAN) which is crucial in marking and maintaining the active centromere throughout the cell cycle [6, 7]. The kinetochore, on the other hand, is important in providing an interface for spindle microtubule binding, stabilizing correct attachments and participating in the spindle assembly checkpoint (SAC), as well as the movement of sister chromatids towards opposite poles during anaphase [8].

Together, the function of the centromere and kinetochore is to ensure high fidelity of chromosome segregation during cell division because an erroneous chromosome segregation can lead to cell arrest or cell death, or more dangerously, chromosomal instability (CIN) and aneuploidy in the daughter cells. CIN, the rate of karyotypic change resulting in anomalous organization and/or number of chromosomes, has been reported as one of the key features in cancer cells and was postulated to precede aneuploidy. Aneuploidy, however, is the karyotypic state depicted by abnormal number of chromosomes and has long been associated with carcinogenesis and birth disorders. The first suggestion of a possible link between aneuploidy and cancer was in the monograph published by Boveri in 1914 [9–11].

Health

The main functional role of the centromere is to ensure that replicated chromosomes are distributed equally to daughter cells during cell division. These functions can be divided into the following classes: (1) Genetic/epigenetic marking or identity of the locus along a specified region of each chromosome, (2) SAC and correct attachment of microtubules, (3) sister chromatid cohesion and release, (4) movement of chromosomes to opposing poles and (5) cytokinesis where a group of transient proteins mark the site for the final separation of the daughter cells.

Centromere Structure

The centromere is comprised of three main zones (Fig. 9.1); (1) the cohesion zone that holds the two replicated sister chromatids together until the onset of anaphase, (2) the DNA interface where proteins directly interact with centromere DNA to mark and maintain the centromere site between cell divisions, and (3) a platform for the capture of spindle microtubules—this zone is commonly referred to as the kinetochore.

Centromere DNA

Human centromere DNA is composed of a tandemly repeated AT-rich monomer of 171 bp commonly known as alpha satellite [12]. This repeat is organized into higher order repeats (HORs) ranging in size from 2 to 35 monomers, which are then organized into further tandem arrays spanning (250 kb to 3 Mb) (Fig. 9.2) [13]. One feature of alpha satellite HORs is that they have chromosome specificity [4]. Differences in the primary sequence of each HOR monomer repeat give rise to its unique chromosome specificity. This difference allows researchers and diagnostic scientists to use techniques such as FISH to identify single chromosomes such as the X or the Y. However, not all chromosomes can be distinguished by a single alpha

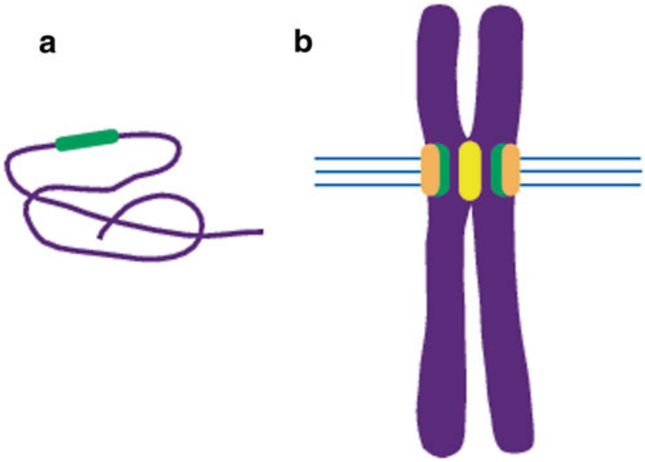


Fig. 9.1 Centromere structure during interphase and metaphase. (**a**) The centromere locus is marked by a group of proteins (*green*) known as the CCAN complex which are found at the same chromosomal site throughout the cell cycle. (b) After DNA replication, the chromatin (*purple*) condenses to form the mature metaphase chromosome attached to spindle microtubules (*blue*). It is during this stage that the centromere attracts other proteins involved in microtubule spindle attachment (*orange*) and sister centromere cohesion (*yellow*)

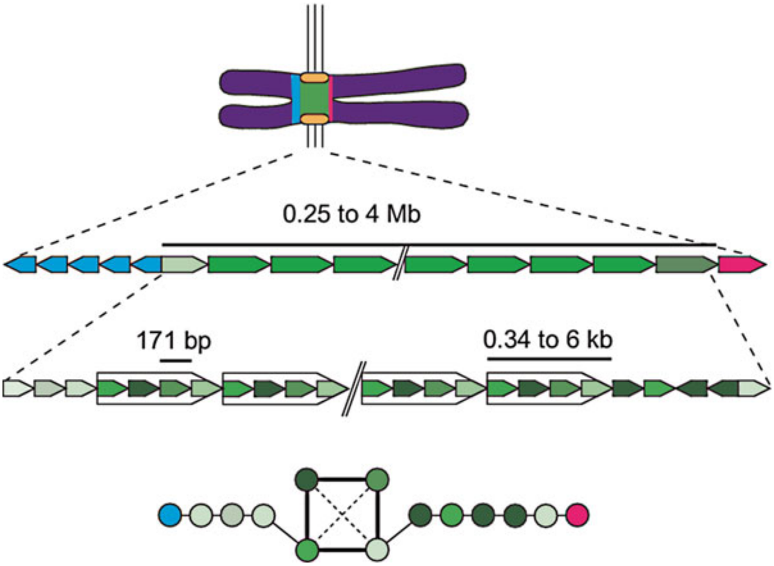


Fig. 9.2 Schematic illustrating the genomic organization of human centromeres. A condensed metaphase chromosome showing centromeric alpha satellite (*green*) and pericentric DNA families (*blue* and *pink*). Note that the repetitive pericentromeric DNA is composed of different sized monomer sequences that are not similar to alpha satellite. The fundamental repeating unit of alpha satellite is an AT-rich 171 bp unit. This is organized into higher order repeats (HORs) that have a high level of sequence identity. Individual centromeres can be distinguished by their HOR type which varies in length and sequence structure. Furthermore, the overall length of the alpha satellite domain is highly polymorphic between individuals. An alternative way to view the organization of centromeric DNA is to break it up into monomer units shown as *colored circles*. The sequential order of each unit is linked with a line and the HOR monomer units are linked with thicker lines which represent multiple HORs

satellite HOR class. This is nicely illustrated in the acrocentric chromosomes, 13, 14, 15, 21 and 22. HORs from these chromosomes share a high level of sequence homology, where single chromosomes cannot be differentiated by hybridization techniques such as FISH or Southern blot.

Alpha satellite DNA also contains a conserved 17-bp motif known as the CENP-B box. This sequence is present in varying frequencies in alpha satellite HORs, ranging from 50 to 0 % in chromosomes 21 and Y, respectively [14]. The CENP-B protein binds to the CENP-B box and is thought to be important for the de novo assembly of the centromere. The formation and stability of artificial chromosomes in the laboratory is dependent on CENP-B-box rich DNA [15]. Paradoxically, once the chromosome is in the cell it does not need the CENP-B protein for full centromere function, as shown by several lines of evidence—(1) knockouts of the CENP-B gene in mouse exhibit full centromere function, develop normally and are fertile, (2) natural centromeres such as in the Y chromosome do not contain the CENP-B box, and (3) human neocentromeres form on DNA that has no alpha satellite or CENP-B box motifs [14, 16–19].

Alpha Satellite DNA Mapping and Sequencing

Alpha satellite DNA was one of the first sequences to be identified and sequenced, however when one examines the centromeric regions of the human reference genome it is quite apparent that they contain megabase-size gaps due to difficulties with contig assembly. While high-throughput genome sequencing has led to a revolution in rapidly identifying the molecular defects behind many human disorders, centromeres remain incomplete due to the short read length of the current parallel sequencing technologies, satellite DNA regions again suffer from poor assembly. Recently, novel computational methods involving unit monomer analysis has provided new ways in analyzing these regions. By grouping similar monomer units together predictions can be made that reveal the overall array length for haploid centromeres such as in the X and Y chromosomes can be made (Fig. 9.2) [20].

CENP-A and Alpha Satellite: Centromeric Chromatin

The centromere-specific histone H3 variant CENP-A (described in "Centromere Proteins" section below) is normally present within subsections of the HOR region of alpha satellite DNA. This has been shown with elegant anti-CENP-A and alpha satellite FISH experiments on extended chromatin fibers [21]. This subdomain structure of CENP-A and alpha satellite is considered to play a role in the three dimensional assembly of the mature mitotic centromere, since alpha satellite DNA is present within the inner (pairing) and outer (microtubule binding) regions of the centromere (Fig. 9.1).

Eviction of the Invaders

Unlike centromeres of multi-cellular eukaryotes, human centromeres are mostly made up of one class of DNA, alpha satellite DNA. It is rare to find the presence of LINE and SINE transposable elements (TEs) within the HOR array. What is the possible mechanism that keeps the intruders at bay? Detailed sequence map analysis at the border regions of alpha satellite has shown that the age of TE insertion decreases as one goes from outer non-alpha satellite DNA to the inner higher order alpha arrays [22]. This suggests that TEs are rapidly pushed away from the HOR region to the periphery. A simple mechanism that would explain this would be unequal crossing over between homologous chromosomes or sister chromatids, which also contributes to the evolution of centromere DNA.

Centromere Proteins

Classes of Centromere/Kinetochore Proteins

The centromere/kinetochore complex can be broadly classified into two main groups based on structure and function with respect to chromosome segregation. The centromere locus needs to be identified and maintained in one region per chromosome, this memory between cell cycles is maintained by a group of proteins that are present at the centromere throughout the cell cycle (Fig. 9.1). The second group falls into the active process of preparing and executing chromosome segregation. These proteins are present at the centromere/kinetochore in a transient manner beginning after DNA replication to the completion of telophase.

To date, over 100 proteins have been shown to locate to the centromere/kinetochore at some stage during the cell cycle. For the purpose of this chapter we are only including proteins that have multiple lines of evidence such as antibody and epitope fusion localization. Some proteins have been misclassified because of artifact signals from antibody staining experiments.

The first set of human centromere proteins discovered were identified using auto-immune sera from patients with scleroderma disease [5]. Protein immunoblotting uncovered three common antigens, named CENP-A, -B and -C in ascending molecular weight order [23]. Serendipitously, these three proteins bind to the centromere DNA and form the foundation platform onto which other centromere and kinetochore proteins assemble the mature, functional structure.

Centromere Function

Epigenetic Marking

Most eukaryotic centromeres are characterized by long tracts of repeat DNA, either satellite or transposable elements. Furthermore, this DNA was often specific to the centromeric locus, for example alpha satellite in humans. One popular hypothesis regarding the interaction between centromere DNA and protein was that the protein had specific DNA-binding affinity, such as the CENP-B protein binding to the CENP-B box motif in alpha satellite [14]. However, immuno-fluorescence analysis of variant chromosomes such as dicentrics or neocentromeres (described in “Disease” section below) showed that some centromere proteins were only present at functionally active centromeres whether alpha satellite DNA was present or absent, and other proteins were present at both active and inactive centromeres [19, 24]. This line of evidence showed that centromeres had genetic and epigenetic characteristics unlike their telomere counterparts which are strictly genetic.

CENP-A: The Primary Mark

CENP-A is a histone H3 variant that is only found at active centromeres [25]. It replaces both units of histone H3 of the histone octamer which provides the centromeric epigenetic mark and a chromatin platform onto which the constitutive centromere-associated network (CCAN) of proteins bind to [6] (see Table 9.1). Further evidence to support the foundation role of CENP-A is shown in gene knockout/knockdown studies which result in the loss of downstream centromere proteins and the absence of a functional kinetochore [26].

Spindle Assembly Checkpoint (SAC)

After chromosomes have replicated and condensed, they then are captured by the mitotic spindle via the interaction with the kinetochore. The chromatid pairs then shuffle between spindle poles to ensure that each sister centromere has attached to spindle microtubules emanating from one pole and thus achieving bi-orientated attachment. Once all chromosomes have acquired correct attachment and equal tension, the chromatids are then ready to segregate to opposite poles. The cell is able to detect the tension and signal for the beginning of anaphase. A group of proteins that are essential for the correct attachment of chromosomes were identified through elegant genetic screens in budding yeast [27, 28]. These spindle assembly checkpoint (SAC) proteins are conserved in humans and mutations elevate the rate of chromosome segregation errors and have a role in cancer predisposition (see "Disease" section).

Sister Centromere Cohesion

After DNA replication, sister chromatids need to be held together to prevent them from prematurely separating, which can result in mis-segregation. A conserved protein complex, known as cohesin, holds the sister chromatids together until the early stages of mitosis when cohesin is progressively removed from the arms and remains at the centromere region until the onset of anaphase. A protector protein, Shugoshin, binds to the centromeric pool of cohesin and thus prevents its premature removal [29]. So the last chromosomal region to be held together before anaphase is the inner centromere domain. In addition to mitosis, cohesin also plays an important role during meiosis I when homologous chromosomes are held together at the centromere by a meiotic-specific cohesin complex. It is hypothesized that weakening of this complex due to aging may contribute to higher rates of chromosomal non-disjunction in women of advanced maternal age (see "Disease" section).

Table 9.1 Centromere and kinetochore proteins organized into functional classes

Centromere protein complex	Proteins	Function
Constitutive Centromere-Associated Network (CCAN)	CENPA, CENPC, CENPH, CENPI, CENPK, CENPL, CENPM, CENPN, CENPO, CENPP, CENPQ, CENPR, CENPS, CENPT, CENPU, CENPW, CENPX	CCAN complex plays a central role in establishing a foundation for mitotic-specific kinetochore proteins.
MIS18 complex	MIS18A, MIS18B, KNL2, PLK1, HJURP	Licensing and loading of CENPA to centromeric chromatin.
KMN network	MIS12, KNL1, DSN1, PMF1, NSL1, NDC80, NUF2, SPC24, SPC25	Outer kinetochore complex required for correct chromosome alignment, mitotic checkpoint signalling and attachment of the kinetochore to microtubules.
RZZ complex	ROD, ZWILCH, ZW10, ZWINT	Mitotic checkpoint role, prevents cells from prematurely exiting mitosis
SKA complex	SKA1, SKA2, SKA3	Microtubule-binding complex required for correct chromosome segregation
Mitotic Checkpoint Complex	BUB1, BUB1B, MAD2L1, CDC20, BUB3, MAD1L1, TTK, AURKB	Ensures all chromosomes have bi-orient attachments to the mitotic spindle, corrects mis-alignment errors and signals onset of anaphase
Chromosome Passenger Complex	INCENP, SURVIVIN, AURKB, BOREALIN	Key regulating complex of mitosis, corrects chromosome mis-alignments, required for chromatin-induced microtubule stabilisation and marks the spindle midzone during anaphase.
Centromere Cohesion	CBX5, CBX1, CBX3, SGOL1, SGOL2, REC8	Holds sister centromeres together around pericentric heterochromatin, protects from premature chromatid separation.
Microtubule-Binding Proteins	CENPE, CENPF, CLASP1, CLASP2, CLIP1, DYNEIN, DYNACTIN, EB1, KIF18A, MCAK, PINX1	Microtubule motor and tracking proteins essential for movement and alignment of chromosomes during mitosis.

Most proteins in this list have been shown to localize to the centromere at some stage during the cell cycle and have a functional role in chromosome segregation. Many other proteins have been found at the centromere but have not been included in this table because they either do not have a clear role in chromosome segregation or their centromeric localization is secondary to their primary function

Chromosome Movement

One of the key roles of the kinetochore is to capture the spindle microtubules, align the chromosomes to the midzone and then move them to the opposite poles. Affinity biochemical experiments from yeast have shown that the budding yeast centromere comprises of one super-complex that binds to one microtubule. Humans contain around 20 microtubules per kinetochore attachment, thus there are multiple subunits that act together in concert [30]. Once each sister centromere is captured to the microtubules they then go through a pushing and pulling action between spindle poles to establish equal tension. This movement is partly triggered by motor, microtubule binding and checkpoint proteins. A protein complex at the heart of this process is the KMN network (Table 9.1). Again, like other complexes, it is conserved in a multitude of eukaryotic organisms and plays an essential role in chromosome segregation. Components of this complex are transiently present at the centromere and form a link between the centromeric chromatin and the outer kinetochore.

Cytokinesis

As described above, centromere cohesion plays an important role in holding the sister chromatids together until the beginning of anaphase. Additional roles include tension sensing and chromosome alignment or error correction. The complex of proteins at the heart of this region is the Chromosome Passenger Complex (CPC) [31]. This includes the four subunits, Aurora B kinase, INCENP, SURVIVIN and BOREALIN. Furthermore, the CPC has an additional role once chromosomes begin to move to opposite poles. They are left behind at the spindle midzone thus marking this region as the site of cellular/cytoplasmic constriction and eventual cleavage of the membranes and spindle microtubules to release the two daughter cells. Any defects in this later stage of mitosis can lead to cells with multiple copies of the genome (polyploidy) and are thought to be involved in tumour progression.

Disease

Structural abnormalities implicating the centromeric DNA, namely the presence of more than one centromere, repositioning of the centromere to a non-centromeric DNA site, prematurely separated centromeres, mutations and aberrant expression of centromere-associated kinetochore proteins, anomalous methylation and altered transcription of alpha satellite, as well as pericentric regions have all been associated with human diseases.

Chromosome Structural Abnormalities

Dicentric Chromosomes

Robertsonian translocations (ROBs) are the most common constitutional structural rearrangements in humans, observed at a rate of one in every thousand live births. ROBs involve whole-arm exchanges between two of the five non-homologous human acrocentric chromosomes (13, 14, 15, 21 and 22), giving rise to a karyotypically metacentric chromosome [32]. Carriers of balanced ROBs are generally normal but with increased risk of infertility due to conception of non-viable fetuses and also with elevated chance of having offspring with Down syndrome.

The other commonly reported constitutional dicentric chromosomes are the isodicentric X chromosomes especially idic(X)(p11) which could occur as both mosaic or non-mosaic. Idic(X)(p11) cases account for about 18 % of Turner syndrome patients, amounting to an incidence rate of approximately 1 in 14,000 females. Other dicentric X chromosomes might include rearranged derivatives of X chromosomes or isodicentrics that have breakpoints at sites other than Xp11 [33].

A rarer non-homologous, non-ROBs had also been reported to give rise to constitutional dicentric chromosomes. Thus far, only 27 cases were reported since the 1970s. Most cases (23/27) involved an acrocentric chromosome and 15/19 of cytogenetically distinguishable heterodicentric chromosomes had only one primary constriction whereby 12/15 of the inactivated centromere being the acrocentric centromeres. This is probably due to the relative stability of the dicentric formed as p-arm deletion of acrocentric chromosomes is not embryonic lethal and the centromeres of acrocentrics have higher tendency to become inactivated [34, 35].

Constitutional dicentric chromosomes are stably transmitted through cell divisions because one of the two centromeres is either inactivated via epigenetic mechanisms or deleted partially or fully (Fig. 9.3) [36, 37]. An inactivated centromere is positive for CENP-B but negative for the essential proteins, CENP-A, -C and -E and hence, is distinguishable from functionally active centromeres [38]. Stability of a dicentric chromosome with two functional centromeres could also be achieved through close proximity of the centromeres—an intercentromeric distance of less than 12 Mb as seen on isodicentric X chromosomes [39].

In malignancies, dicentric chromosomes are generally an outcome of telomere fusion events due to telomere instability of cancer cells as observed in giant cell tumor of the bone, meningioma, chronic lymphocytic leukemia (CLL), pancreatic cancer and osteosarcoma [40, 41]. However, most dicentric chromosomes in hematological malignancies arise from reciprocal translocation that produces a dicentric chromosome and an acentric chromosomal fragment which might be lost in subsequent mitoses. Thus far, the mechanism of centromere inactivation in malignancies has not been well studied. Investigations into the dicentric chromosomes of acute myeloid leukemia (AML) and myelodysplastic syndromes indicated that a repertoire of strategies namely functional (epigenetic) inactivation, intercentromeric deletion, inversion to reduce intercentromeric distance, and partial or full centromere excision were deployed to produce a more stable chromosome [41].

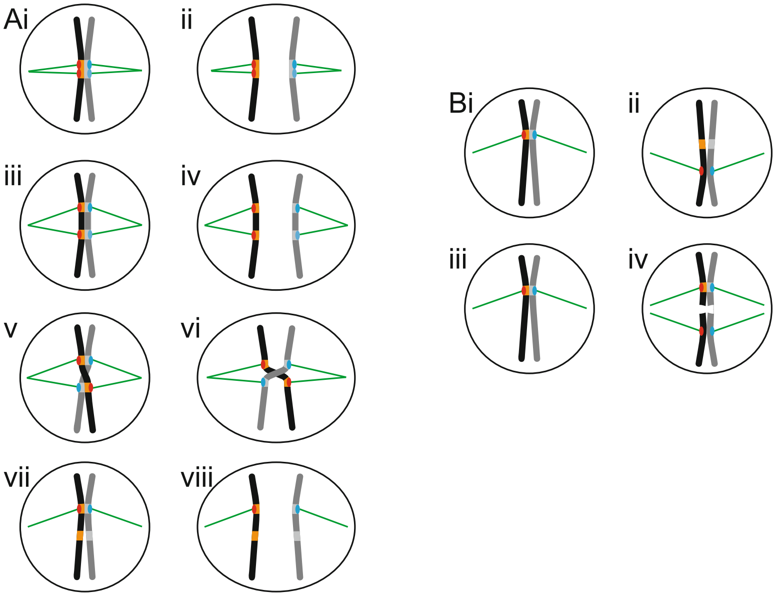


Fig. 9.3 Epigenetic status of the centromere in abnormal chromosomes. Replicated sister chromatids (*black* and *grey*) are shown aligned and attached to microtubules. The satellite-rich centromere DNA (*orange* and *light grey* shaded boxes) mark the centromere locus. Functionally active centromeres build a mature kinetochore (*red* and *blue ovals*) which capture spindle microtubules and move chromatids to opposite poles. (*Ai* and *ii*) Functional dicentric chromosomes with closely spaced centromeres act in unison to correctly segregate the chromatids. (*Aiii* and *iv*) Dicentric chromosomes with centromeres spaced further apart can also segregate correctly but (*Av* and *vi*) sister chromatids can twist between the two centromeres resulting in single chromatids attached to both poles which causes possible breakage of the chromosome. (*Avii* and *viii*) Epigenetic inactivation of one of the centromeres (loss of the kinetochore) resolves the conflict between the two active centromeres and thus chromosomes can correctly segregate. Neocentromeres form on non-alpha satellite DNA, often in euchromatic regions. (*B*) Two possible mechanisms of neocentromere formation, (*Bi* and *ii*) repositioning of the centromere to a new region along the chromosome. The old centromere is subsequently inactivated. (*Biii* and *iv*) Another mechanism shows a breakage and the formation of an acentric fragment. This chromosomal fragment is rescued by the formation of a kinetochore but the underlying alpha satellite DNA is absent

Neocentromeres

Neocentromere is the term coined for an ectopic centromere which forms in a region of the chromosome outside the repetitive alpha satellite DNA [19]. It binds all known centromere proteins except CENP-B and functions similarly to the native centromere [42] although the level of CENP-A incorporation [43], cohesion [44, 45] and error correction by Aurora B [46] appear to be lowered. Neocentromeres have been found in euchromatic sites and the formation of neocentromeres does not seem to correlate with reduced expression of the genes in those regions [47].

The first report of a constitutional human neocentromere in 1993 was from cytogenetic screening of a 4 year-old patient who was presented with delayed speech development [19]. Subsequent discoveries were made in patients with a wide spectrum of clinical presentations including facial dysmorphism and growth retardation in younger patients to infertility and high proportion of miscarriages in adult patients [47]. In children, several cancer types including retinoblastoma [48], Wilms tumor [49], cystic hygroma [50] and hemangioma [51] were reported as co-morbidities with the other developmental disorders.

In addition, neocentromeres have also been specifically associated with a few cancers thus far, namely AML, atypical lipomas and well-differentiated liposarcomas (ALP-WDLPS), lung sarcomatoid carcinoma and T-cell non-Hodgkin lymphoma [52]. The presence of neocentromeres on either a supernumerary ring or a long marker chromosome, both derived from the long arm of chromosome 12, is a defining characteristic of ALP-WDLPS of borderline malignancy [53]. These chromosomes have amplification of the 12q14-15 region containing oncogenes that include *MDM2* and *CDK4* [54]. However, the same amplified region is also found in other more aggressive liposarcomas but on chromosome 12 with alpha satellites suggesting that the neocentromere formed was to stabilize the complex rearranged acentric chromosome containing amplified 12q14-15 which might confer selective advantage within the tumor microenvironment besides highlighting the difference between neocentromere and the native centromere with alpha satellites [47].

Premature Centromere Division

Premature centromere division (PCD; OMIM #212790) is a cytogenetically detectable trait where the X chromosome appears to have no discernible centromere resulting in a rod-shaped X chromosome. The frequency of lymphocytes showing PCD and DNA damage increases as we age but for sporadic Alzheimer's disease patients, the increased frequency was even more significant when compared to their age-matched controls. In addition, PCD was shown to be consistently more prominent in females than males and was thought to be the cause of chromosomal instability resulting in tissue mosaicism and neuronal cell death in Alzheimer's disease [55].

PCD is also found in older females who experience significantly higher chance of spontaneous abortion and bearing children with trisomies especially trisomy 21. In females, the immature oocytes arrest in prophase I and only proceed with meiosis upon hormonal stimulation during the period after puberty until menopause. Hence, the chiasmata between homologous chromosomes and cohesion of the sister chromatid arms in prophase I as well as the subsequent centromere cohesion between the sister chromatids in meiosis II have to be properly maintained by the cohesin complexes for many years before these oocytes are released and potentially fertilized [56]. This long period of arrest led to the postulation of an age-dependent 'cohesin fatigue' being a contributing factor to the much higher aneuploidy rate of oocytes in older women [57, 58].

Centromere Protein Genes

CENP-A and HJURP in Cancer

Studies performed in colorectal, testicular, liver, breast and lung cancers were reported to have elevated expression of CENP-A while separate studies in lung, breast and brain cancers had reported on overexpression of the CENP-A chaperone, HJURP [59, 60]. CENP-A and HJURP could potentially be used as prognostic markers for certain groups of cancer. CENP-A has been demonstrated to correlate positively with pathological grade and negatively with survival prognosis in lung adenocarcinoma [61], epithelial ovarian cancer [62] and estrogen-receptor positive breast cancers that were not treated with systemic therapy [63]. HJURP has shown a similar pattern of correlation with astrocytomas, the most common type of adult brain cancer [59]. In combination, upregulation of both CENP-A and HJURP at their mRNA levels were found to be associated with decreased survival in breast cancer patients [64].

BUB1B, ESCO2, CASC5 and CENP-E in Developmental Disorders

Mosaic variegated aneuploidy syndrome (MVA; OMIM #257300) is a collective term for the cytogenetic characteristic where mosaic aneuploidies are commonly observed with clinical features namely microcephaly, mental retardation and growth retardation. In a subset of MVA patients, premature chromatid separation (PCS; OMIM #176430) was evident [65]. PCS is another cytogenetic description for a spectrum of diseases, in which a significant percentage of the mitotic lymphocytes appear to have separated centromeres and splayed chromatids. This is in contrast to the metaphase chromosome of normal, colchicine or colcemid treated cells where two sister chromatids are linked at the centromere region [66].

The SAC gene, *BUB1B* was not only the first gene found to be associated with MVA but also the first mitotic SAC gene where its allelic mutations in the germline were linked to a human disease [67]. Monoallelic *BUB1B* mutations appeared to give rise to the most severe phenotype including high occurrence of PCS, cataracts, Dandy–Walker syndrome and cancer. Biallelic *BUB1B* mutations yielded moderate phenotype while MVA without *BUB1B* mutations rarely had PCS and showed no signs of cataracts, Dandy–Walker syndrome and cancer [68]. Hence, many have postulated that other mitotic SAC genes might have important role in instigating the remaining forms of MVA.

The other cytogenetically observed trait around the centromere is heterochromatin repulsion which is most noticeable on chromosomes with large tracts of heterochromatin namely chromosomes 1, 9 and 16 [69]. This affects most metaphase chromosomes of patients with Roberts syndrome (RBS; OMIM #268300) and the milder SC phocomelia syndrome (SC; OMIM #269000). The causative gene for both of these syndromes was found to be Establishment of Cohesion 1 Homologue 2 (*ESCO2*) and these syndromes can be regarded as a spectrum depending on the variants of the mutated *ESCO2*. Clinical features of these patients include growth retardation, mental retardation and the presence of craniofacial abnormalities with microcephaly being the most common besides several others including hypertelorism, hypoplastic nasal alae and malar hypoplasia. The presence of cleft lip and palate was associated with the severity of limbs malformations while corneal opacities correlated with mental retardation and cardiac defects [70].

CASC5 or *KNL1* mutations were reported to cause autosomal recessive primary microcephaly (MCPH; OMIM #251200). CASC5 is a member of the conserved KMN (KNL1/MIS12 complex/NDC80 complex) network of proteins within a kinetochore that links the chromosome to the microtubules. CASC5 which localizes to the kinetochore from G2 til late anaphase is also part of the SAC machinery as it is known to bind to BUB1B [71]. Compound heterozygous variants of *CENPE* had recently been described in two siblings with microcephalic osteodysplastic primordial dwarfism (MOPD2; OMIM #210720) which was a disease previously reported to be linked to mutated centrosome-associated protein, pericentrin (PCNT) [72]. CENP-E is a dimeric kinesin-like motor protein which was shown to be important for the stability of binding between kinetochore and the dynamic microtubules, while PCNT is essential in the formation of microtubule arrays at the centrosome [73]. This suggests that the overlapping phenotype for both *CENPE* and *PCNT* mutations might be spindle-related [72].

Other Centromere Protein Genes in Cancers

Kinetochore protein genes that are crucial for the normal function of the centromere have been reported to be mutated or differentially expressed in various cancers. Mutations in *BUB1* were implicated in colorectal, adult T-cell leukemia/lymphoma (ATLL), lung and pancreatic cancers while mutations in *BUB1B* had been reported in more cancer types including colorectal cancer, MVA, ATLL, glioblastomas, Wilms tumor and B-cell lymphoma [74].

In addition to mutation, the level of expression for SAC proteins appears to be important in tumorigenesis. *BUB1*, *BUB1B* and *BUB3* were reported to be unregulated in gastric cancer [75]. However, in pediatric glioblastoma, expression of *BUB1* and *BUB1B* were upregulated whereas *BUB3* was downregulated [76]. In clear cell renal carcinomas investigated for the expression of their SAC genes, *BUB1*, *BUB1B* and *MAD2L1* (MAD2 mitotic arrest deficient-like 1) were found to be overexpressed while *MAD1* had decreased expression [77].

Epigenetics

Epigenetics is the study of the changes in gene expression or protein function that are not due to alterations in the DNA sequence of the gene, but are heritable through cell division. Such changes could occur at, (1) the genome structural level involving

DNA methylation, histone modifications, nucleosome positioning, and histone variants, (2) the RNA level which includes RNA splicing and RNA interference, and (3) the protein level in cases of prion formation where the 'infectious' proteins are able to induce conformational change of native proteins rendering them 'infectious' as well as dysfunctional [78, 79].

DNA Methylation

Cytosine residues of the non-CpG islands (non-CGIs) or else referred to as the global CpG dinucleotides within intronic and intergenic regions, especially transposable elements and simple repeat sequences, are mostly methylated in somatic tissues as opposed to unmethylated cytosines in the CGIs that are known to coincide with gene promoters or regulatory regions [80].

Abnormal DNA methylation had been reported in immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome. ICF is a rare autosomal recessive disease which is currently categorized into three groups namely ICF1 (OMIM #242860) with mutations found in the DNA methyltransferase 3B (*DNMT3B*) gene, ICF2 (OMIM #614069) with mutations in zinc finger- and BTB domain-containing 24 (*ZBTB24*) gene, and the final group with currently unknown molecular etiology provisionally designated ICFX [81]. Although all three groups exhibit hypomethylation of satellites 2 and 3 which are part of the constitutive heterochromatin, ICF2 and ICFX, however, show additional hypomethylation at the alpha satellite [82]. In the heterochromatic region that exhibit reduced DNA methylation from an average level of 80 % in normal cells to 30 % in ICF cells, some heterochromatic genes were shown to have escaped silencing compared to the control, although each patient appeared to have his own signature of heterochromatic genes that escaped silencing across different chromosomes [83].

Wilms tumor is the most common renal tumor in children under 5 years of age, accounting for 90 % of the total pediatric renal cancer cases and contributing to approximately 7 % of all pediatric malignancies [84]. Hypomethylation of alpha satellite on chromosomes 1 and 10 was observed in Wilms tumor patient samples but it was not correlated with aneuploidy. To a lesser extent and frequency, satellite 2 was also hypomethylated on these chromosomes [85]. These studies into ICF and Wilms tumors indeed pose an interesting question about the mechanisms that lead to the differences in their hypomethylation profiles.

In cancer as well as aged cells, global hypomethylation and concomitant increase in the methylation of promoters have been observed and were thought to contribute to genomic instability and gene silencing respectively. Furthermore, global non-CGIs could be further subcategorized and studied. Hypomethylation of Alu, LINE-1 and alpha satellite in CLL patients were examined and alpha satellite hypomethylation was suggested to be a potential negative prognostic marker for CLL [86]. Separately, in a study performed in ovarian epithelial tumors, satellite 2 hypomethylation on chromosomes 1 and 16 was strongly correlated with both genome-wide hypomethylation and the degree of tumor malignancy. Extensive hypomethylation of chromosome 1 alpha satellite was also observed in larger proportion of carcinomas compared to the more benign forms [80]. Therefore, it appears that the level of DNA methylation at centromeric satellite sequences is a useful biomarker in the study and diagnosis of cancers.

Pericentric and Centromeric Transcription and Histone Modifications

Expression of pericentric and centromeric transcripts has been found to be altered in senescent cells and human cancer samples when compared to normal tissues, reflecting the global epigenetic deregulation [87]. This could partly be facilitated by the altered DNA methylation in these regions. In addition to global DNA methylation changes, the global histone marks have also been found to be altered. The loss of both H4K20me3, the pericentric constitutive heterochromatin mark, and H3K27me3, the facultative heterochromatin mark, were reported in lung cancer cells when examined with the non-tumor cells [88].

Upregulation of pericentric satellite 3 was also observed in a Hutchinson-Gilford progeria syndrome (HGPS) patient. HGPS (OMIM #176670) is a disease of rapid aging due to the expression of mutant Lamin A, a developmentally regulated gene. However, the expression of alpha satellite was unaltered, suggesting that the expression of pericentric and centromeric sequences are controlled by different mechanisms. The upregulation of satellite 3 was accompanied by the loss of H3K27me3 and pericentric constitutive heterochromatin mark H3K9me3 but by the increase of another constitutive heterochromatin mark, H4K20me3 [89]. Hence, thus far, the relationships between the expression of repetitive satellites and both DNA methylation as well as histone modifications remain to be clarified.

Although the cases aforementioned were characterized by upregulation of centromeric and/or pericentromeric sequences, the right transcriptional balance between sense and antisense strand of both pericentric and centromeric sequences appears to be crucial for the proper formation and function of a centromere [90, 91].

Evolution

Primate Centromere DNA

Alpha satellite DNA is a relatively conserved centromeric repeat family. It is found in great apes, old world monkeys and new world monkeys, which span approximately 43 million years of evolution since the last common ancestor. In great apes it is organized into HOR structures, however in more distant species, it is mainly found in divergent monomeric forms. One proposed hypothesis is that HOR structure arose after the divergence of the great apes from the rest of the primate species

[92]. Interestingly, HOR structures are not necessarily restricted to certain species or particular chromosomes, for example the centromere array of the mouse Y chromosome contains a HOR but the autosomes and X chromosome only have the monomeric form [93].

As described above, alpha satellite is found in many primate species across 43 million years of evolution. This appears to be rather a long time when compared to other mammalian centromere satellites such as the mouse minor satellite DNA. The monomer repeat unit is 120 bp and is only found in a subset of species in the *Mus* genus spanning about 5–7 million years [94]. The higher rate of centromere DNA evolution in the mouse may be related to a much shorter generation time which increases the chance of the centromere array to rearrange and diverge during meiotic recombination. Even though minor and alpha satellite DNAs are quite diverged, they do share some features such as a high AT content and the conservation of the CENP-B box motif.

The Rapidly Evolving Y Centromere

The human Y chromosome exists in a haploid state in males, and offers a unique perspective into the evolution of centromere DNA within a species since there is no homologous counterpart of this region for meiotic recombination to occur. Like most other Y chromosome sequences that do not recombine with a homologue, the centromere DNA has undergone a rapid rate of sequence divergence. Some of the features that mark the Y centromere as separate from other human centromeres include; a diverged alpha satellite monomer, absence of the CENP-B box motif, diverged HOR and a significantly smaller overall length [13, 14]. Evidence for the rapid divergence in the Y alpha satellite sequence is nicely illustrated in the analysis of the HOR in humans and chimpanzees. The HOR of the X and 17 alpha satellite exhibits a conserved co-linearity of the HOR, whereas the Y alpha satellite has completely lost this conserved structure and the length of the HOR between humans and chimpanzees is also different [93].

The functional consequences of a rapidly diverging and smaller Y centromere may be responsible for the Y chromosome's partial instability during division of aging cells [95, 96]. Measurement of the CENP-A protein on Y centromeres shows that it contains around half the amount when compared to the autosomes and X centromeres [43]. This lower amount is consistent with less alpha satellite DNA present at Y centromeres. Y alpha satellite DNA ranges in length from 250 to 1500 kb, in contrast to the X centromere which is megabases in size, ranging from 1300 to 3700 kb [13]. On the extreme end of low amounts of alpha satellite, neocentromeres are formed on non-alpha satellite regions of the genome and are found to contain even lower amounts of CENP-A than the Y centromere [43] (see "Disease" section).

Adaptive Evolution of Foundation Centromere Proteins

It has been hypothesized that rapidly evolving centromere DNA can expand and create larger centromeres that can bind more spindle microtubules [97]. Other evolutionary mechanisms can increase the size of a centromere, such as translocations of acrocentric chromosomes in humans to generate metacentric (Robertsonian) chromosomes with two adjacent centromeres [98]. These chromosomes have a higher chance of being inherited during the asymmetric cell divisions of female meiosis where the egg spindle pole releases more microtubules to capture the bigger centromere than the polar body pole. To prevent a complete runaway of chromosomes with larger and larger centromeres the cell counters this expansion by epigenetic means, through the adaptive evolution of centromere chromatin proteins such as CENP-A and CENP-C [99]. Evidence for this hypothesis is now accumulating in many species groups, such as flies, plants and primates that show these two proteins are under adaptive evolution [100–102]. In contrast, when similar sequence analysis across the primate group was performed, it did not show any evidence for adaptive selection for the non-essential CENP-B protein, even though this protein directly binds to the alpha satellite DNA [102].

ZNF397, an Evolutionary New Centromere Protein

Many centromere proteins are conserved in eukaryotic species, ranging from the single-celled budding yeast to humans. In some instances in evolution, new proteins appear via a variety of molecular mechanisms. One example of this is zinc finger protein 397 (ZNF397), which presumably arose from a gene duplication event after the separation of placental and marsupial mammals. We had previously identified ZNF397 using anti-centromere antibodies from a patient with autoimmune disease [103]. Interestingly, this protein has a unique cell cycle localization pattern where it is present from the end of telophase through to early prophase. Knockout experiments in mouse showed that the protein is not essential for chromosome segregation. One attractive hypothesis is that the protein has acquired centromere targeting activity but it is yet to be directly involved in full kinetochore function.

Karyotype Evolution and Meiotic Drive: Robertsonian Translocations

Mendel's law of segregation implies that the two homologous chromosomes in a parent segregate at meiosis into the gametes to ensure the offspring acquire only one copy of each chromosome from each parent, thereby maintaining the proper chromosome number in sexually reproducing organisms. This law assumes that the process of segregation occurs in a random and non-biased manner. However, in humans, the most common ROBs namely der(13q14q) and der(14q21q) arise mainly during oogenesis but not spermatogenesis [32]. It was postulated that the chromosomal rearrangements of ROBs cause functional heterozygosity at the centromere of homologous chromosomes leading to the differential interactions with the meiotic spindle. This then contributes to preferential segregation of the rearranged chromosomes into functional meiotic product instead of the polar bodies [104]. These ROB chromosomes in male carriers are not subjected to the same meiotic drive owing to the process of spermatogenesis where polar bodies do not form, hence, do not render the opportunity for preferential segregation.

Neocentromeres and Evolutionary New Centromeres

One of the hallmarks of speciation at the genetic level is the divergence of karyotypes between two newly-formed species. This separation can often produce a reproductive barrier between the two groups which then further accelerates the rate of evolution. Chromosomal rearrangements including, translocations, inversions, deletions and duplications, are a driving force in the emergence of new karyotype configurations. As a consequence of genomic rearrangements, the centromere can also change position to either rescue an acentric chromosomal fragment or compensate for a partially deleted (inactivated) centromere, as has been observed in de novo clinical cytogenetic cases (Fig. 9.3) (see “Disease” section). Changes in centromere position in closely related species led to the concept of Evolutionary New Centromeres (ENCs) [105]. ENCs were initially thought to have arisen due to the physical repositioning of an extant centromere. This hypothesis has been replaced by the ENC hypothesis because of more accurate genome and cytogenetic mapping. So the evolutionary timeline of the formation of an ENC is as follows: (1) chromosomal rearrangement, (2) neocentromere formation and (3) accumulation of satellite DNA at the neocentromeric locus [106].

A good example that illustrates this progress from neocentromere to an ENC is in the orangutan. Early cytogenetic analyses of orangutan centromeres using alpha satellite FISH showed that at least one chromosome was devoid of alpha satellite DNA [107]. It wasn’t until high-throughput sequencing and CENP-A pulldown technologies that definitively revealed that chromosome 12 contained a neocentromere but had not yet acquired any alpha satellite DNA sequences [108]. The ENC chromosome 12 is present in 2 species of orangutan, *Pongo abelii* (Sumatran) and *Pongo pygmaeus* (Bornean) which shared a common ancestor between 0.4 and 1 million years ago. This shows that ENCs can take a long time to acquire satellite sequences. Interestingly, the progenitor chromosome 12 with the alphoid centromere still exists together with the ENC form in the two orangutan species.

Conclusions

In the last few decades we have made rapid progress in the discovery of most of the genomic and protein elements that make up a functional centromere. Human centromere DNAs have been identified and mapped to each chromosome. Current sequencing methods have made some in-roads towards completing the genome map of these repeat-rich regions. Furthermore, novel computational methods have allowed the interrogation of high-throughput genome sequencing results from individuals, however, gaps still remain. The next breakthrough in long-read single-molecule sequencing will allow these gaps to be closed and analyzed centromere-by-centromere. Insights will be made in the rate of evolution across populations and within families. This will enable researchers to further understand the contribution of variation and mutation on centromere dysfunction in human chromosome instability disorders.

References

1. Battaglia E (2003) Centromere, kinetochore, kinochore, kinetosome, kinosome, kinetomere, kinomere, kinetocentre, kinocentre: history, etymology and intepretation. Caryologia 56:1–21
2. Arrighi FE, Hsu TC (1971) Localization of heterochromatin in human chromosomes. Cytogenetics 10:81–86
3. Vissel B, Choo KH (1992) Evolutionary relationships of multiple alpha satellite subfamilies in the centromeres of human chromosomes 13, 14, and 21. J Mol Evol 35:137–146
4. Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. Nucleic Acids Res 19:1179–1182
5. Moroi Y, Peebles C, Fritzler MJ, Steigerwald J, Tan EM (1980) Autoantibody to centromere (kinetochore) in scleroderma sera. Proc Natl Acad Sci U S A 77:1627–1631
6. Hori T, Shang W-H, Takeuchi K, Fukagawa T (2013) The CCAN recruits CENP-A to the centromere and forms the structural core for kinetochore assembly. J Cell Biol 200:45–60
7. Westhorpe FG, Straight AF (2013) Functions of the centromere and kinetochore in chromosome segregation. Curr Opin Cell Biol 25:334–340
8. Foley EA, Kapoor TM (2013) Microtubule attachment and spindle assembly checkpoint signalling at the kinetochore. Nat Rev Mol Cell Biol 14:25–37
9. Bayani J, Selvarajah S, Maire G, Vukovic B, Al-Romaih K, Zielenska M et al (2007) Genomic mechanisms and measurement of structural and numerical instability in cancer cells. Semin Cancer Biol 17:5–18
10. Chandhok NS, Pellman D (2009) A little CIN may cost a lot: revisiting aneuploidy and cancer. Curr Opin Genet Dev 19:74–81
11. Holland AJ, Cleveland DW (2009) Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat Rev Mol Cell Biol 10:478–487
12. Manuelidis L (1978) Chromosomal localization of complex and simple repeated human DNAs. Chromosoma 66:23–32
13. Wevrick R, Willard HF (1989) Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. Proc Natl Acad Sci U S A 86:9394–9398

14. Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol 109:1963–1973
15. Okada T, Ohzeki J, Nakano M, Yoda K, Brinkley WR, Larionov V et al (2007) CENP-B controls centromere formation depending on the chromatin context. Cell 131:1287–1300
16. Hudson DF, Fowler KJ, Earle E, Saffery R, Kalitsis P, Trowell H et al (1998) Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. J Cell Biol 141:309–319
17. Perez-Castro AV, Shamanski FL, Meneses JJ, Lovato TL, Vogel KG, Moyzis RK et al (1998) Centromeric protein B null mice are viable with no apparent abnormalities. Dev Biol 201:135–143
18. Kapoor M, Montes de Oca Luna R, Liu G, Lozano G, Cummings C, Mancini M et al (1998) The cenpB gene is not essential in mice. Chromosoma 107:570–576
19. Voullaire LE, Slater HR, Petrovic V, Choo KH (1993) A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? Am J Hum Genet 52:1153–1163
20. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ (2014) Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res 24:697–707
21. Blower MD, Sullivan BA, Karpen GH (2002) Conserved organization of centromeric chromatin in flies and humans. Dev Cell 2:319–330
22. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. Science 294:109–115
23. Earnshaw WC, Rothfield N (1985) Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma 91:313–321
24. Earnshaw WC, Migeon BR (1985) Three related centromere proteins are absent from the inactive centromere of a stable isodicentric chromosome. Chromosoma 92:290–296
25. Palmer DK, O'Day K, Wener MH, Andrews BS, Margolis RL (1987) A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. J Cell Biol 104:805–815
26. Howman EV, Fowler KJ, Newson AJ, Redward S, MacDonald AC, Kalitsis P et al (2000) Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. Proc Natl Acad Sci U S A 97:1148–1153
27. Hoyt MA, Totis L, Roberts BT (1991) S. cerevisiae genes required for cell cycle arrest in response to loss of microtubule function. Cell 66:507–517
28. Li R, Murray AW (1991) Feedback control of mitosis in budding yeast. Cell 66:519–531
29. Salic A, Waters JC, Mitchison TJ (2004) Vertebrate shugoshin links sister centromere cohesion and kinetochore microtubule stability in mitosis. Cell 118:567–578
30. Santaguida S, Musacchio A (2009) The life and miracles of kinetochores. EMBO J 28:2511–2531
31. Carmena M, Wheelock M, Funabiki H, Earnshaw WC (2012) The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis. Nat Rev Mol Cell Biol 13:789–803
32. Bandyopadhyay R, Heller A, Knox-DuBois C, McCaskill C, Berend SA, Page SL et al (2002) Parental origin and timing of de novo Robertsonian translocation formation. Am J Hum Genet 71:1456–1462
33. Scott SA, Cohen N, Brandt T, Warburton PE, Edelmann L (2010) Large inverted repeats within Xp11.2 are present at the breakpoints of isodicentric X chromosomes in Turner syndrome. Hum Mol Genet 19:3383–3393
34. Lemyre E, der Kaloustian VM, Duncan AM (2001) Stable non-Robertsonian dicentric chromosomes: four new cases and a review. J Med Genet 38:76–79
35. Dutta UR, Pidugu VK, Dalal A (2012) Molecular cytogenetic characterization of a non-Robertsonian dicentric chromosome 14;19 identified in a girl with short stature and amenorrhea. Case Rep Genet 2012:212065

36. Page SL, Shaffer LG (1998) Chromosome stability is maintained by short intercentromeric distance in functionally dicentric human Robertsonian translocations. Chromosome Res 6:115–122
37. Rivera H, Ayala-Madrigal LM, Gutiérrez-Angulo M, Vasquez AI, Ramos AL (2003) Isodicentric Y chromosomes and secondary microchromosomes. Genet Couns 14:227–231
38. Page SL, Earnshaw WC, Choo KH, Shaffer LG (1995) Further evidence that CENP-C is a necessary component of active centromeres: studies of a dic(X; 15) with simultaneous immunofluorescence and FISH. Hum Mol Genet 4:289–294
39. Sullivan BA, Willard HF (1998) Stable dicentric X chromosomes with two functional centromeres. Nat Genet 20:227–228
40. Gisselsson D, Jonson T, Petersén A, Strömbeck B, Dal Cin P, Höglund M et al (2001) Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. Proc Natl Acad Sci U S A 98: 12683–12688
41. Mackinnon RN, Campbell LJ (2011) The role of dicentric chromosome formation and secondary centromere deletion in the evolution of myeloid malignancy. Genet Res Int 2011:643628
42. Saffery R, Irvine DV, Griffiths B, Kalitsis P, Wordeman L, Choo KH (2000) Human centromeres and neocentromeres show identical distribution patterns of >20 functionally important kinetochore-associated proteins. Hum Mol Genet 9:175–185
43. Irvine DV, Amor DJ, Perry J, Sirvent N, Pedeutour F, Choo KHA et al (2004) Chromosome size and origin as determinants of the level of CENP-A incorporation into human centromeres. Chromosome Res 12:805–815
44. Amor DJ, Bentley K, Ryan J, Perry J, Wong L, Slater H et al (2004) Human centromere repositioning “in progress”. Proc Natl Acad Sci U S A 101:6542–6547
45. Alonso A, Hasson D, Cheung F, Warburton PE (2010) A paucity of heterochromatin at functional human neocentromeres. Epigenetics Chromatin 3:6
46. Bassett EA, Wood S, Salimian KJ, Ajith S, Foltz DR, Black BE (2010) Epigenetic centromere specification directs aurora B accumulation but is insufficient to efficiently correct mitotic errors. J Cell Biol 190:177–185
47. Marshall OJ, Chueh AC, Wong LH, Choo KHA (2008) Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am J Hum Genet 82:261–282
48. Morrissette JD, Celle L, Owens NL, Shields CL, Zackai EH, Spinner NB (2001) Boy with bilateral retinoblastoma due to an unusual ring chromosome 13 with activation of a latent centromere. Am J Med Genet 99:21–28
49. Hu J, McPherson E, Surti U, Hasegawa SL, Gunawardena S, Gollin SM (2002) Tetrasomy 15q25.3 → qter resulting from an analphoid supernumerary marker chromosome in a patient with multiple anomalies and bilateral Wilms tumors. Am J Med Genet 113:82–88
50. Haddad V, Aboura A, Tosca L, Guediche N, Mas A-E, L’Herminé AC et al (2012) Tetrasomy 13q31.1qter due to an inverted duplicated neocentric marker chromosome in a fetus with multiple malformations. Am J Med Genet A 158A:894–900
51. Liu J, Jethva R, Del Vecchio MT, Hauptman JE, Pascasio JM, de Chadarévian J-P (2013) Tetrasomy 13q32.2qter due to an apparent inverted duplicated neocentric marker chromosome in an infant with hemangiomas, failure to thrive, laryngomalacia, and tethered cord. Birth Defects Res A Clin Mol Teratol 97:812–815
52. Burrack LS, Berman J (2012) Neocentromeres and epigenetically inherited features of centromeres. Chromosome Res 20:607–619
53. Gaskin CM, Helms CA (2004) Lipomas, lipoma variants, and well-differentiated liposarcomas (atypical lipomas): results of MRI evaluations of 126 consecutive fatty masses. AJR Am J Roentgenol 182:733–739
54. Italiano A, Maire G, Sirvent N, Nuin PAS, Keslair F, Foa C et al (2009) Variability of origin for the neocentromeric sequences in analphoid supernumerary marker chromosomes of well-differentiated liposarcomas. Cancer Lett 273:323–330

55. Zivković L, Spremo-Potparević B, Siedlak SL, Perry G, Plećaš-Solarović B, Milićević Z et al (2013) DNA damage in Alzheimer disease lymphocytes and its relation to premature centromere division. Neurodegener Dis 12:156–163
56. Nagaoka SI, Hassold TJ, Hunt PA (2012) Human aneuploidy: mechanisms and new insights into an age-old problem. Nat Rev Genet 13:493–504
57. Jessberger R (2012) Age-related aneuploidy through cohesion exhaustion. EMBO Rep 13:539–546
58. Wassmann K (2013) Sister chromatid segregation in meiosis II: deprotection through phosphorylation. Cell Cycle 12:1352–1359
59. Valente V, Serafim RB, de Oliveira LC, Adorni FS, Torrieri R, Tirapelli DP et al (2013) Modulation of HJURP (Holliday Junction-Recognizing Protein) levels is correlated with glioblastoma cells survival. PLoS One 8:e62200
60. Vardabasso C, Hasson D, Ratnakumar K, Chung C-Y, Duarte LF, Bernstein E (2014) Histone variants: emerging players in cancer biology. Cell Mol Life Sci 71:379–404
61. Wu Q, Qian Y-M, Zhao X-L, Wang S-M, Feng X-J, Chen X-F et al (2012) Expression and prognostic significance of centromere protein A in human lung adenocarcinoma. Lung Cancer 77:407–414
62. Qiu J-J, Guo J-J, Lv T-J, Jin H-Y, Ding J-X, Feng W-W et al (2013) Prognostic value of centromere protein-A expression in patients with epithelial ovarian cancer. Tumour Biol 34:2971–2975
63. McGovern SL, Qi Y, Pusztai L, Symmans WF, Buchholz TA (2012) Centromere protein-A, an essential centromere protein, is a prognostic marker for relapse in estrogen receptor-positive breast cancer. Breast Cancer Res 14:R72
64. Hu Z, Huang G, Sadanandam A, Gu S, Lenburg ME, Pai M et al (2010) The expression level of HJURP has an independent prognostic impact and predicts the sensitivity to radiotherapy in breast cancer. Breast Cancer Res 12:R18
65. Callier P, Faivre L, Cusin V, Marle N, Thauvin-Robinet C, Sandre D et al (2005) Microcephaly is not mandatory for the diagnosis of mosaic variegated aneuploidy syndrome. Am J Med Genet A 137:204–207
66. Kajii T, Ikeuchi T (2004) Premature chromatid separation (PCS) vs. premature centromere division (PCD). Am J Med Genet A 126A:433–434
67. Hanks S, Coleman K, Reid S, Plaja A, Firth H, Fitzpatrick D et al (2004) Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. Nat Genet 36:1159–1161
68. García-Castillo H, Vásquez-Velásquez AI, Rivera H, Barros-Núñez P (2008) Clinical and genetic heterogeneity in patients with mosaic variegated aneuploidy: delineation of clinical subtypes. Am J Med Genet A 146A:1687–1695
69. Goh ES-Y, Li C, Horsburgh S, Kasai Y, Kolomietz E, Morel CF (2010) The Roberts syndrome/SC phocomelia spectrum--a case report of an adult with review of the literature. Am J Med Genet A 152A:472–478
70. Vega H, Trainer AH, Gordillo M, Crosier M, Kayserili H, Skovby F et al (2010) Phenotypic variability in 49 cases of ESCO2 mutations, including novel missense and codon deletion in the acetyltransferase domain, correlates with ESCO2 expression and establishes the clinical criteria for Roberts syndrome. J Med Genet 47:30–37
71. Genin A, Desir J, Lambert N, Biervliet M, Van Der Aa N, Pierquin G et al (2012) Kinetochore KMN network gene CASC5 mutated in primary microcephaly. Hum Mol Genet 21:5306–5317
72. Mirzaa GM, Vitre B, Carpenter G, Abramowicz I, Gleeson JG, Paciorkowski AR et al (2014) Mutations in CENPE define a novel kinetochore-centromeric mechanism for microcephalic primordial dwarfism. Hum Genet 133:1023–1039
73. Gudimchuk N, Vitre B, Kim Y, Kiyatkin A, Cleveland DW, Ataullakhanov FI et al (2013) Kinetochore kinesin CENP-E is a processive bi-directional tracker of dynamic microtubule tips. Nat Cell Biol 15:1079–1088

74. Bolanos-Garcia VM, Blundell TL (2011) BUB1 and BUBR1: multifaceted kinases of the cell cycle. Trends Biochem Sci 36:141–150
75. Grabsch H, Takeno S, Parsons WJ, Pomjanski N, Boecking A, Gabbert HE et al (2003) Overexpression of the mitotic checkpoint genes BUB1, BUBR1, and BUB3 in gastric cancer--association with tumour cell proliferation. J Pathol 200:16–22
76. Morales AG, Pezuk JA, Brassesco MS, de Oliveira JC, de Paula Queiroz RG, Machado HR et al (2013) BUB1 and BUBR1 inhibition decreases proliferation and colony formation, and enhances radiation sensitivity in pediatric glioblastoma cells. Childs Nerv Syst 29: 2241–2248
77. Pinto M, Vieira J, Ribeiro FR, Soares MJ, Henrique R, Oliveira J et al (2008) Overexpression of the mitotic checkpoint genes BUB1 and BUBR1 is associated with genomic complexity in clear cell kidney carcinomas. Cell Oncol 30:389–395
78. Halfmann R, Lindquist S (2010) Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits. Science 330:629–632
79. Grossniklaus U, Kelly WG, Kelly B, Ferguson-Smith AC, Pembrey M, Lindquist S (2013) Transgenerational epigenetic inheritance: how important is it? Nat Rev Genet 14:228–235
80. Qu G, Dubeau L, Narayan A, Yu MC, Ehrlich M (1999) Satellite DNA hypomethylation vs. overall genomic hypomethylation in ovarian epithelial tumors of different malignant potential. Mutat Res 423:91–101
81. Weemaes CMR, van Tol MJD, Wang J, van Ostaijen-ten Dam MM, van Eggermond MCJA, Thijssen PE et al (2013) Heterogeneous clinical presentation in ICF syndrome: correlation with underlying gene defects. Eur J Hum Genet 21:1219–1225
82. Jiang YL, Rigolet M, Bourc'his D, Nigon F, Bokesoy I, Fryns JP et al (2005) DNMT3B mutations and DNA methylation defect define two types of ICF syndrome. Hum Mutat 25:56–63
83. Brun M-E, Lana E, Rivals I, Lefranc G, Sarda P, Claustres M et al (2011) Heterochromatic genes undergo epigenetic changes and escape silencing in immunodeficiency, centromeric instability, facial anomalies (ICF) syndrome. PLoS One 6:e19464
84. Fawkner-Corbett DW, Howell L, Pizer BL, Dominici C, McDowell HP, Losty PD (2014) Wilms' tumor--lessons and outcomes--a 25-year single center UK experience. Pediatr Hematol Oncol 31:400–408
85. Ehrlich M, Hopkins NE, Jiang G, Dome JS, Yu MC, Woods CB et al (2003) Satellite DNA hypomethylation in karyotyped Wilms tumors. Cancer Genet Cytogenet 141:97–105
86. Fabris S, Bollati V, Agnelli L, Morabito F, Motta V, Cutrona G et al (2011) Biological and clinical relevance of quantitative global methylation of repetitive DNA sequences in chronic lymphocytic leukemia. Epigenetics 6:188–194
87. Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ et al (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science 331:593–596
88. Eymery A, Horard B, El Atifi-Borel M, Fourel G, Berger F, Vitte A-L et al (2009) A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. Nucleic Acids Res 37:6340–6354
89. Shumaker DK, Dechat T, Kohlmaier A, Adam SA, Bozovsky MR, Erdos MR et al (2006) Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. Proc Natl Acad Sci U S A 103:8703–8708
90. Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KHA et al (2012) Active transcription and essential role of RNA polymerase II at the centromere during mitosis. Proc Natl Acad Sci U S A 109:1979–1984
91. Hall LE, Mitchell SE, O'Neill RJ (2012) Pericentric and centromeric transcription: a perfect balance required. Chromosome Res 20:535–546
92. Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE (2007) Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. PLoS Comput Biol 3:1807–1818

93. Pertile MD, Graham AN, Choo KHA, Kalitsis P (2009) Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. Genome Res 19:2202–2213
94. Garagna S, Redi CA, Capanna E, Andayani N, Alfano RM, Doi P et al (1993) Genome distribution, chromosomal allocation, and organization of the major and minor satellite DNAs in 11 species and subspecies of the genus Mus. Cytogenet Cell Genet 64:247–255
95. Griffin DK, Abruzzo MA, Millie EA, Sheean LA, Feingold E, Sherman SL et al (1995) Nondisjunction in human sperm: evidence for an effect of increasing paternal age. Hum Mol Genet 4:2227–2232
96. Nath J, Tucker JD, Hando JC (1995) Y chromosome aneuploidy, micronuclei, kinetochores and aging in men. Chromosoma 103:725–731
97. Zwick ME, Salstrom JL, Langley CH (1999) Genetic variation in rates of nondisjunction: association of two naturally occurring polymorphisms in the chromokinesin nod with increased rates of nondisjunction in Drosophila melanogaster. Genetics 152:1605–1614
98. Daniel A (2002) Distortion of female meiotic segregation and reduced male fertility in human Robertsonian translocations: consistent with the centromere model of co-evolving centromere DNA/centromeric histone (CENP-A). Am J Med Genet 111:450–452
99. Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293:1098–1102
100. Malik HS, Henikoff S (2001) Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics 157:1293–1298
101. Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S (2002) Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. Plant Cell 14:1053–1066
102. Schueler MG, Swanson W, Thomas PJ, NISC Comparative Sequencing Program, Green ED (2010) Adaptive evolution of foundation kinetochore proteins in primates. Mol Biol Evol 27:1585–1597
103. Bailey SL, Chang SC, Griffiths B, Graham AN, Saffery R, Earle E et al (2008) ZNF397, a new class of interphase to early prophase-specific, SCAN-zinc-finger, mammalian centromere protein. Chromosoma 117:367–380
104. Pardo-Manuel de Villena F, Sapienza C (2001) Nonrandom segregation during meiosis: the unfairness of females. Mamm Genome 12:331–339
105. Rocchi M, Stanyon R, Archidiacono N (2009) Evolutionary new centromeres in primates. Prog Mol Subcell Biol 48:103–152
106. Kalitsis P, Choo KHA (2012) The evolutionary life cycle of the resilient centromere. Chromosoma 121:327–340
107. Miller DA, Sharma V, Mitchell AR (1988) A human-derived probe, p82H, hybridizes to the centromeres of gorilla, chimpanzee, and orangutan. Chromosoma 96:270–274
108. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM et al (2011) Comparative and demographic analysis of orang-utan genomes. Nature 469:529–533

Chapter 10
Structure and Functions of Telomeres in Organismal Homeostasis and Disease

Penelope Kroustallaki and Sarantis Gagos

Telomeres Co-evolved with Linear Chromosomes

From the origins of life, the ancestral circular prokaryotic genome developed a complex biological machinery committed to maintain DNA integrity that ensures cellular homeostasis and the fidelity of the transmission of precise genetic information through cell generations [1–3]. In both prokaryotes and eukaryotes, a highly efficient repair system, evolved to sense and restore DNA base pair alterations or miss-matches, as well as any discontinuity of the genome, such as single or double strand breaks of the DNA helix [4, 5]. DNA repair is a highly dynamic biological process that implies the congression of several types of specialized factors at the site of the DNA lesion, and in parallel, interacts with the course of the progression of the cell cycle affecting dramatically the overall cellular fate [6, 7]. Activation of the DNA repair machinery is operated by a series of post-translational modifications of numerous protein substrates that are proficient to sense the site of the damage, they are capable to slow down the cell cycle as long as the damage is repaired, and they are even destined to kill the cell if the damage is unrepairable [8].

In eukaryotes, the circular ancestral genome evolved into multiple linear chromosomal entities that in many species can be visualized by light microscopy in cytological preparations stained by nucleophilic dyes [9, 10]. The adaptivity of the ancestral linear chromosomes was challenged by two very important biological constrains that literally shaped life as we know it: (a) The so-called, "chromosome end-replication problem" proposed independently by Watson and Olovnikov during the 1970s [11–13] and (b) the putative activation of DNA damage responses at the

P. Kroustallaki, B.Sc., M.Sc. • S. Gagos, Ph.D. (✉)
Laboratory of Genetics, Department of Experimental Medicine and Translational Research,
Biomedical Research Foundation of the Academy of Athens, Greece (BRFAA),
4 Soranou Ephessiou St., Athens 115 27, Greece
e-mail: penelopekroust@hotmail.com; sgagos@bioacademy.gr

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_10

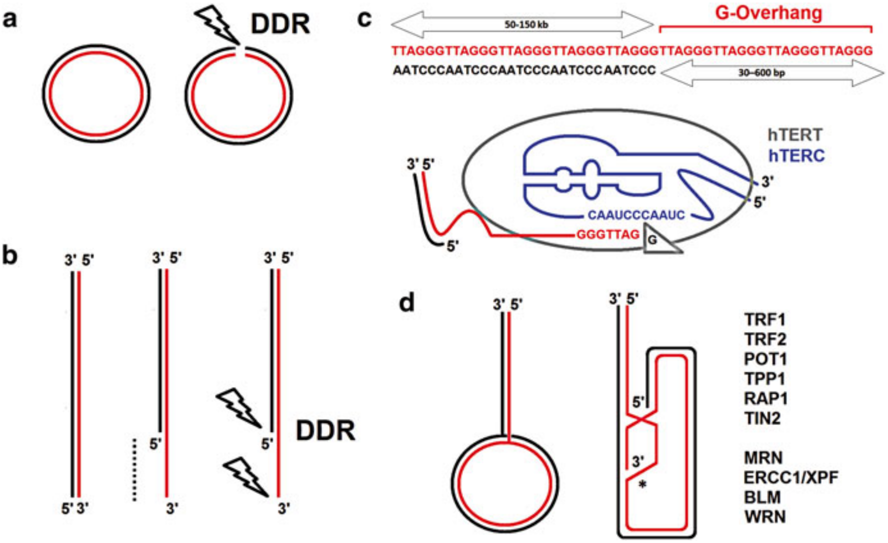


Fig. 10.1 The two major telomeric constrains and nature's solutions to the chromosome terminal problem: Any interruption of the integrity of the prokaryotic circular DNA molecule in the form of DSB, activates DNA damage responses (DDR) that may lead to cell cycle arrest, or even to cell death (**a**). The evolution of circular chromosomes to linear DNA molecules was challenged by the universal insufficiency of the DNA replication machinery to replicate completely the 5′-end of the linear DNA molecule. This process shortens the chromosome by each cell division and can jeopardize genome integrity and organismal homeostasis. In addition, the termini of linear DNA molecules can be perceived as DNA lesions, they can activate local DDR and can be degraded by nucleases (**b**). The partially dispensable repetitive, G-rich, primary structure of eukaryotic telomeres and the development of telomere replenishment mechanisms, such as the ribozyme telomerase, resolved the end-replication problem. Like most telomerases, human telomerase reverse transcriptase (hTERT) assembles with its RNA component (hTERC), anchors itself to the telomere, and neo-synthesizes G-rich telomeric repeats according to its single-stranded complementary RNA template (**c**). To resolve the second telomeric constrain, the ends of linear chromosomes evolved secondary structures that mimic the circular ancestral DNA and can hide the single-stranded telomeric G-overhang into a telomeric loop (T-loop). The T-loop is formed by the invasion of the G-overhang into double stranded telomeric repeats that is mediated by a three-stranded DNA displacement loop (D-loop) (*asterisk*). The formation of T-loops is facilitated by shelterin components and telomere interacting factors (**d**)

linear chromosome's "blunt" ends [14, 15] or the catastrophic activity of DNA exonucleases at the exposed chromosome termini [16] (Fig. 10.1a, b).

The "chromosome end-replication problem" stems from the inert insufficiency of the semiconservative DNA replication machinery that is incapable to fully polymerize the lagging DNA strand and eventually trims down a terminal part from the 5′ end of the linear chromosome every time a DNA molecule replicates itself and the cell divides [17]. The second telomeric constrain, entailed the immediate evolution of a specialized terminal chromosomal structure that hides the discontinuity of the DNA helix from the DDR machinery and in parallel, protects chromosome blunt ends from nucleolytic degradation [18–20].

The specialized entities that comprise linear chromosome ends, were termed "telomeres" from the Greek words "τέλος" (-end) and "–μέρος" (part) by the Nobel laureate geneticist Herman Muller in 1938 [21]. The role of telomeres in chromosome protection and overall genome integrity was recognized during the 1930s by another Nobel laureate, Barbara McClintock who performed a series of studies in irradiated maize (*Zea mays*) cells and described the ongoing mutagenic phenomenon of Breakage/Fusion/Bridge cycles (B/F/B cycles) [22, 23]. Dividing cells undergoing subsequent rounds of terminal chromatid, or chromosome breakage, followed by fusion and anaphase bridging, display multiple structural chromosome anomalies such as dicentric (or poly-centric) and ring chromosomes [23, 24]. The B/F/B cycles are capable to generate chromosomal translocations, inversions or duplications and can lead to losses of large genomic segments or even whole chromosomes [25–28]. Today, it is well established that B/F/B cycles are illegitimate outcomes of Non-Homologous End Joining (NHEJ) or Homologous Recombination (HR) DNA repair pathways and play an important role in the ongoing chromosomal instability that is considered a hallmark of cancer [29, 30].

The Primary Nucleotide Structure of Telomeres and Mechanisms of Telomere Length Maintenance

The molecular era of telomere research begins in 1978, when E. Blackburn and J. Gall, described that the terminal structures of the bulky extrachromosomal rDNA molecules of *Tetrahymena*, are composed by tandemly repeated hexanucleotide sequences [31]. In the early 1980s, J. Szostak and E. Blackburn, showed that protozoan telomeric function from *Tetrahymena*, could be transferred to yeast chromosomes of *Saccharomyces cerevisiae* [32]. Driven by these findings, E. Blackburn, and C. Greider, described for the first time, that eukaryotic telomeres consist of tandemly repeated G- and C-rich complementary DNA sequences and set the basis to resolve the end-replication problem [33]. Their subsequent studies revealed that in *Tetrahymena*, the G-rich telomeric DNA strand can be neo-synthesized by an RNA-dependent, reverse transcriptase, termed "telomerase" [33, 34]. These discoveries immediately raised the interest of the broad scientific community that very soon called telomeres the "thread of life" and telomerase the "fountain of youth". The important contributions of E. Blackburn, J. Szostak, and C. Greider to science were recognized by the 2009 Nobel Prize in Physiology or Medicine.

In the past three decades, our understanding about the structure and the function of telomeres has significantly improved through the efforts of a relatively small, but vigorously thriving research community: In 1989, Morin [35] showed that the primary structure of human telomeres is composed by tandem repeats of the complementary hexa-nucleotides TTAGGG/AATCCC (G-rich and C-rich strands). The same group documented human telomerase activity, while J. Meyne used Fluorescent in Situ Hybridization (FISH) to visualize telomeres at chromosome termini, showing in parallel that the TTAGGG repeats are highly conserved between mammals [35, 36]. In fact, from anthozoans (corals) to humans, different species share the same TTAGGG repeat at their chromosome termini indicating that telomere homeostasis is operated by highly conserved biological pathways [20, 37–39].

The primary nucleotide structure of mammalian telomeres is composed of long stretches of double-stranded TTAGGG/AATCCC hexa-nucleotide repeats that vary in size between different organisms and species [36, 40]. The double telomeric strand terminates in a relatively short, single-stranded, TTAGGG-rich overhang that protrudes out of the 3′-end of each linear DNA molecule [41]. The length of the G-overhang ranges between 30 and 600 nucleotides [36, 41, 42]. The average length of double-stranded mammalian telomeres varies between chromosomes of the same cell and amongst species, from 5–20 kb in humans, to 50–150 kb in several rodents [36, 43, 44]. The telomerase holoenzyme, acts as a reverse transcriptase to polymerize the 3′-overhang of the telomeric sequence using as template a specialized nuclear non-coding RNA that contains sequence complementarity with the G-rich telomeric repeat [34] (Fig. 10.1c).

Studies in *Caenorhabditis elegans*, revealed that double stranded telomeres can also terminate in 5′-end single stranded C-rich overhangs [45]. Similar to the worms, G1/S phase arrested or terminally differentiated mammalian cells, display frequent C-overhangs [46]. Telomeric C-rich overhangs are less well studied than G-overhangs and may play a role in biological processes involving homologous telomeric recombination [46].

Replicative Senescence and Ageing

Normal human cells *in culture* have a limited proliferative life-span, often termed the “Hayflick limit”. After 50–100 population doublings (PDs), most of the cultured cells enter a static phase termed “senescence”. Senescent cells stop to divide, undergo “crisis” and eventually they die [47, 48]. The term “replicative senescence”, was introduced by Greider and Harley in 1990, who elaborated the “telomeric hypothesis of aging”, based on their observations in human diploid cells that progressively lost their telomeres after consequent population doublings (PDs) [49]. According to this theory, most dividing somatic cells do not possess an active mechanism of telomere maintenance; hence, anytime a cell replicates its genome, a portion of the telomeres will be lost and cannot be replenished [11, 13, 49]. Thus gradual telomeric shortening must be associated with cellular, tissue and organismal ageing. Indeed, several studies have shown that there is a positive correlation with telomere length of somatic tissues and life-span [50, 51]. Furthermore, cells from individuals with premature aging syndromes show reduced telomeric length and proliferative capacity *in culture* [52, 53]. However, there is a marked deviation of telomere length between individuals belonging in the same age group, indicating genetic and multifactorial effects on telomere metabolism [54–57].

In longed-lived mammals and humans, the gradual telomeric loss and the loss of replicative capacity act as barriers to carcinogenesis [58, 59]. On the contrary, many short-lived animals such as laboratory mice, have significantly longer telomeres than humans, and do not display replicative aging [44, 60, 61]. There is now substantial evidence that in short lived mammals, like rodents, cellular senescence does not act as an oncosuppressor pathway [62–64]. Primary normal diploid mouse cell cultures undergo spontaneous transformation and become polyploid, upon a limited number of PDs, whereas normal human cells are constrained by the Hayflick limit [64, 65]. Ectopic expression of telomerase activity in cultured, diploid, human retinal pigment epithelial cells is capable to confer immortalization and bypasses the Hayflick barrier [66].

Biogenesis of Telomerases

Although telomerase activity can be reconstituted in vitro just by the combined presence of the reverse transcriptase (TERT) and its RNA template (TERC), the different telomerase holoenzymes between species, are composed from several protein components [67, 68]. These molecules take part in biogenesis, assembly, targeting and regulation of telomerases. The molecular cloning of the telomerase holoenzyme constituents, in various experimental in vivo models and in human cells, enabled the investigation and manipulation of the biological processes that regulate telomere length [60, 67, 69, 70]. In *S. cerevisiae*, the telomerase ribonucleoprotein (RNP) is composed of the three protein subunits, Est1, Est2 and Est3 [71–73]. Depletion of any one of these proteins in vivo, is associated with a yeast phenotype described as “Ever Shorter Telomeres” and leads to massive cell death through replicative senescence [74, 75].

Based on sequence similarity Reichenbach et al., and Snow et al. (2003) identified the Suppressor of Morphogenesis in Genitalia (SMG), hEST1A(SMG6), hEST1B(SMG5) and hEST1C(SMG7) proteins as human homologues of the yeast Est genes. Despite the weak sequence homology, hEST1A and hEST1B were proven to be part of the active human telomerase holoenzyme [76, 77]. The conserved moiety of hEST1C revealed a 14-3-3 phosphoserine binding motif, but no association with telomerase RNP [78]. In yeast and humans, telomerase is considered to act as a multimer (dimer in humans) [79–81] while in protozoa like Tetrahymena, acts as a monomer [82, 83]. Telomerase activity is regulated by negative feed-back mechanisms. In *Saccharomyces Cerevisiae* the telomerase dependent telomere elongation was found increased in chromosomes carrying shorter telomeres [84].

Telomerase RNAs have been identified in several protozoa, yeast and vertebrate species including humans [85]. Despite the extensive phylogenetic divergence in primary and secondary structures [86] the RNA region that bears the template of all telomerases, appears to be always single-stranded and is capable to associate with the active reverse transcriptase site of TERT [86] (Fig. 10.1c). In humans, precursor hTR molecules, are transcribed by RNA polymerase II and immediately associate with a protein heterotrimer composed of Dyskerin, NHP2, and NOP10, that is bound to the RNP assembly chaperone NAF1 [87, 88]. The process is facilitated by two helicases, Pontin (TIP49, TIP49a, or RUVBL1) and Reptin (TIP48, TIP49b, or RUVBL2), and by the Dyskerin chaperone SHQ1 [89–92].

Cajal bodies are subnuclear compartments that accumulate RNP particles involved in splicing, ribosome biogenesis and telomere maintenance [93]. After co-transcriptional assembly with the Dyskerin heterotrimer and RNA processing, hTERC is directed towards the nucleoli and then to Cajal bodies [94–96]. The helicases Pontin and Reptin, bind to Dyskerin and TERT independently and facilitate the assembly of the telomerase RNP [95]. The mature telomerase RNP concentrates in Cajal bodies by the aid of TCAB1/WDR79 (Telomerase Cajal Body protein 1) and then it is rooted to the telomeres [16, 95, 97, 98]. The two subunits of the Ku heterodimer implicated in Non-Homologous End Joining (NHEJ) DNA repair, associate directly to the telomerase RNA component to facilitate telomere neosynthesis [99]. Telomerase holoenzyme recognizes the telomere, and binds its substrate DNA, through sequence complementarity encoded by its RNA template [95]. The di-hexa-nucleotide RNA template (CAAUCCCAAUCC) of hTERC catalyzes the addition of a single nucleotide at the time, to the 3′ end of the telomeric G-overhang [100] (Fig. 10.1c). Neo-synthesis of the next telomeric nucleotide, requires a slight translocation of the active site along the template (type I translocation) [33, 95, 101]. If during this process telomerase remains associated to the telomere, the holoenzyme is considered to present nucleotide processivity [102]. In vitro DNA synthesis by most telomerases is Nucleotide Processive [103]. When the final telomeric nucleotide of the template is synthesized, a realignment of the catalytic site to the beginning of the RNA di-hexa-nucleotide is required, to allow a new round of reverse transcription (translocation II). If telomerase remains bound to the telomere during translocation II, is considered to display Repeat Processivity [33, 101]. Different species exert variable degrees of relative repeat addition processivity of their telomerases [35, 101, 104–106]. In cultured human cancer cell lines, the pharmaceutical disruption of repeat addition processivity inhibits telomerase activity, inducing telomere shortening and replicative senescence [107]. Evaluation of telomerase processivity can be achieved using the original protocol applied for the detection of telomerase activity by Greider and Blackburn [33]. The method takes advantage of the enzymatic activity of telomerase that is capable to elongate terminal repeats in the presence of radio-labelled dNTPs [108], the products of the reaction are visualized as DNA ladders by gel electrophoresis [33].

Measuring Telomere Length

The classical procedure to measure the length of repetitive telomeric sequences is based on extensive digestion of genomic DNA with frequent cutting restriction enzymes, that do not break down telomeric and subtelomeric regions (Terminal Restriction Fragment assay) [49, 109–111]. Size resolution of sub-terminal and terminal restriction fragments is achieved via agarose gel electrophoresis. Telomeric DNA is visualized as smears by Southern blotting or in-gel hybridization, using probes specific for C-or G-rich telomeric repeats, hence it presents extensive length variation between cells and chromosomes [57, 111]. The TRF assay is considered to be the most accurate method to measure the length of telomeric repeats in epidemiological studies [112]. Currently there are several methods to measure the length of telomeric repeats in biological samples, clinical specimens, cell nuclei or even individual chromosome ends [113]. Telomeric in situ Quantitative-FISH or Telomeric Flow-FISH are commonly used in both research and clinical settings [114–119]. These two protocols are based on FISH of telomere specific fluorescent PNA (peptide nucleic acid analog) probes that stoichiometrically bind to terminal DNA repeats and can quantify telomeric length in situ, based on the intensity of the emitted fluorescence signals [120–124].

The arsenal of telomere length quantification protocols has been expanded by the introduction of PCR-based methods such as Single Telomere Length Analysis (STELA) [54, 125] and of MM-Q-PCR that utilizes Real-Time PCR technology [126]. STELA is a DNA ligation-based method that uses primers designed for specific subtelomeric sequences, and can accurately measure telomere length of specific chromosome ends [125]. STELA can be modified to evaluate extremely short telomeres [127, 128]. MM-Q-PCR normalizes the amplification rate of telomeric DNA to a single gene, to validate the overall amount of telomeres in a given biological specimen [126]. MM-Q-PCR requires small amounts of DNA and is extensively applied in large scale population studies [126, 129, 130].

Telomeres and Cancer

In 1990, de Lange et al. [131] showed that primary tumors display shorter telomeres than their adjacent normal tissues and proposed the “telomeric theory of cancer”. De Lange’s hypothesis was supported by the findings of Counter et al. [132], who revealed that telomerase is activated in ovarian tumor cells, but not in stromal nonmalignant cells, suggesting that telomerase activity might be linked to continuous proliferation of cancer cells. The studies on telomerase activity in neoplasia boost up from 1994, when J.W. Shay and W.E. Wright, presented the Telomeric Repeat Amplification Protocol (TRAP) assay to evaluate telomerase activity with high sensitivity. In Kim et al. (1994), the two researchers and their colleagues, used the TRAP technology to demonstrate for the first time, that the majority of human tumors activate telomerase [133, 134]. The TRAP assay has been modified through the years and became quantitative by Q-RT-PCR [108], it’s general principle though, remains the same: In protein extracts, an oligonucleotide with the TTAGGG repeat is incubated with unlabeled nucleotides. Then the products of the reaction are PCR amplified by the addition of a G-rich complementary primer that bears a short tract of non-telomeric DNA at the 5′ end to prevent repetitive sequence miss-alignment [133].

An alternative mechanism of telomere elongation (Alternative Lengthening of Telomeres-ALT) that is mediated by homologous recombination of telomeric repeats was discovered in 1993 in yeast mutants deficient for telomerase activity [74]. Three years later, the group of R. Reddel, presented for the first time, that a small proportion of human immortalized or cancer cell lines do not express telomerase activity, but they utilize an alternative pathway of telomere lengthening (ALT-pathway) [135]. In cancer or virally immortalized human cell lines, the ALT pathway is associated with a “loose” heterochromatin structure at telomeres and centromeres [136, 137], and is characterized by increased homologous telomeric recombination, extreme deviation of telomere length from very short, to as long as 50–60 kb, frequent presence of the so called ALT associated PML bodies, and extensive chromosomal instability [136, 138–145]. In addition, cells utilizing the ALT pathway display abundant extrachromosomal, C-rich, telomeric repeats (C-circles) and increased rates of telomeric sister chromatid exchanges (T-SCEs) [146] (Fig. 10.2c). The ALT-associated PML bodies (APBs) are ALT-characteristic, sub-nuclear compartments that contain Promyelocytic Leukemia body protein (PML), telomeric DNA, telomere associated factors and proteins involved in DDR (i.e. RAD51, RPA1, 53BP1 and the MRN complex (MRE11, RAD50, NBS1) [147–152]. APBs are enriched in G2/M [153, 154] and may play a role in telomere recombination by tethering together chromosome ends and by promoting heterologous telomeric interactions [139, 141, 155]. The orphan nuclear receptors of the NR2C/F classes (TR2, TR4, COUP-TF1, COUP-TF2 and EAR2), which belong to the nuclear hormone receptor (NHR) family of transcription factors, are frequently found at the telomeres of ALT cells and may favor the telomere–telomere recombination necessary for ALT maintenance [156]. There is now increasing evidence that embryonic and somatic stem cells display both known mechanisms of telomere maintenance [157–159].

Mouse Models

Telomerase activity knock-out mice have normal development [160]. Reduced fertility and degenerative defects in highly proliferating tissues emerge only after three generations [160, 161]. The phenotype is more pronounced in the sixth generation, with severe congenital malformations, male sterility and an increase in the incidence of spontaneous lymphomas and carcinomas [160, 161]. In addition, tumor cells from late generation double knockout mice, null for mTERC and p53, showed elevated frequencies of chromosome fusions, anaphase bridges, and nonreciprocal translocations [162]. Conversely, telomerase over-expression in aging transgenic mice was associated with spontaneous emergence of epidermal tumors, mammary and lung carcinomas or lymphomas [163–167]. Hence, both telomerase depletion and overexpression, can lead to carcinogenesis in mice. Lack of telomerase activity generates genomic instability and promotes tumorigenesis [161, 162], while excess of telomerase facilitates neoplastic transformation since most tumors depend on telomerase to maintain continuous cellular proliferation [30, 133].

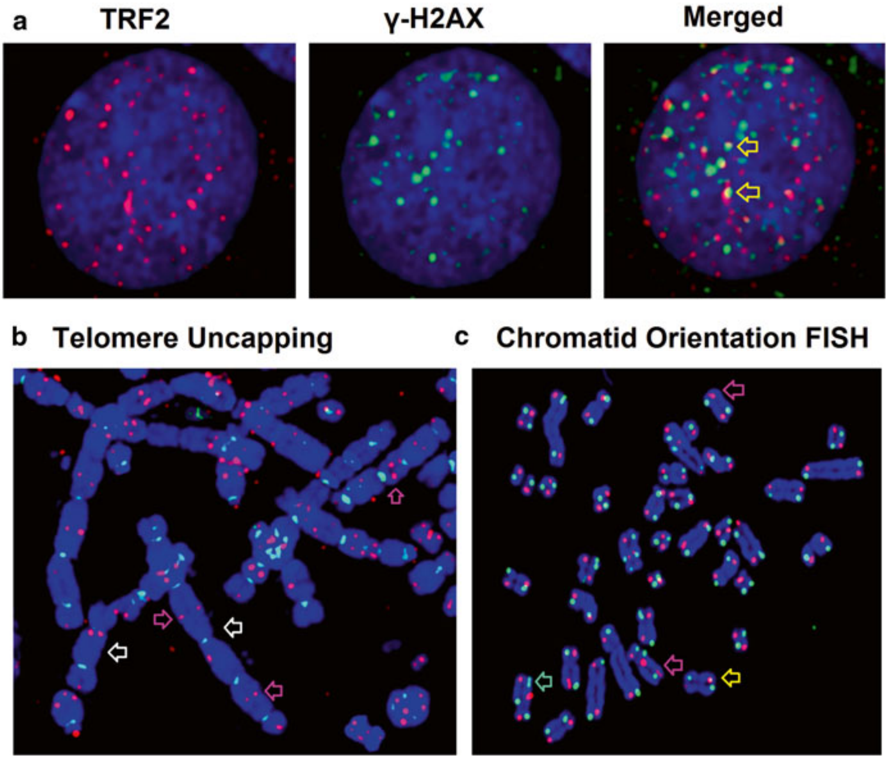


Fig. 10.2 Telomere dysfunction induced foci (TIFs) (*arrows*) in a cell nucleus from an immortalized human cell line (VA-13) 24 h after exposure to gamma-irradiation. Immuno-Fluorescence microscopy reveals co-localizations (*yellow spots*) of the telomere specific protein TRF2 (shelterin component) with the phosphorylated histone H2A (γ-H2AX) that is a marker of DNA damage responses. TRF2 is labelled with Alexa-568 (*red*), γ-H2AX with Alexa-488 (*green*). DAPI (4′,6-diamidino-2-phenylindole) is *blue*, (630×) (**a**). A partial metaphase spread from human BJ fibroblasts depleted for TRF2 and counter-stained with DAPI (630×). Fluorescence In Situ Hybridization (FISH) with probes specific for all human centromeres (labelled green with FITC) and for TTAGGG telomeric repeats (labelled red with Rhodamine), display multi-centromeric chromosomes generated by extensive Non-Homologous End Joining (NHEJ) of uncapped chromosome termini (*white arrows*). Junction points maintain visible telomeric repeats (*red arrows*) (**b**). Strand specific telomeric Chromatid Orientation FISH in pig iPS cells [366] demonstrates Telomeric Sister Chromatid Exchanges (T-SCE) (*yellow arrows*) and induction of fragile telomeres (FT) in both the G-and C-rich telomere strands (*red* and *green arrows*). Peptide Nucleic Acid analog (PNA) probes specific for TTAGGG (green = FITC) and AATCCC repeats (red = Rhodamine), DAPI is *blue* (×630). T-SCEs have been associated with increased telomeric recombinogenicity in the ALT pathway and extreme telomere dysfunction when telomeres are critically shortened [367, 368]. FTs have been connected to increased telomeric replication stress [211]

The first attempt to extend longevity in genetically engineered mice through telomere manipulation was presented by the group of M. Blasco, in 2008, who showed that overexpression of mTERT in cancer resistant Sp53/Sp16/SARF mice, can delay aging [168]. In 2011, Jaskelioff et al. [169] proved that in prematurely aged mTERT deficient mice, reconstitution of telomerase activity can revert a number of age-related phenotypes caused by extreme telomere shortening. Prompted by these findings, in 2012 de Jesus and associates examined the effects of adenoviraly-mediated increase at the levels of TERT, in naturally aged laboratory mice. Interestingly the life-span of these animals was significantly increased with no effects in cancer susceptibility. Therefore, at the organismal level, telomerase acts as a longevity gene by preventing premature telomere attrition [170]. Understanding the complex mechanisms of telomere length regulation, will provide in the near future the means to ameliorate the biologic consequences of replicative ageing and in parallel to protect tissues and organs from neoplasia.

The High Order Structure of Telomeres and Nature's Solution to the Second Basic Telomeric Constrain

The description of the repetitive nature of the telomeric DNA sequences and the biological pathways of telomere elongation, effectively resolved the first telomeric constrain set by the end replication problem [34]. Nature's solution to the second telomeric limitation, required a complex structural organization of the telomeric territories, specialized to efficiently protect chromosome termini from being perceived as targets of DNA damage responses or enzymatic degradation [19, 171]. The peculiar telomere protective structure is shaped both by the unique properties of primary telomeric DNA sequences, and by a dynamic interaction of telomeric repeats with telomere associating nuclear factors [15, 20, 171]. Albeit some sequence dissimilarities between species, most eukaryotic telomeres are rich in Guanines [39] this unique property provides to the telomeric repeats the ability to fold into non-canonical secondary four-stranded DNA structures formed by Guanine-quartets, known as G-quadruplexes [20, 172–174]. These unusual DNA conformations are much more stable than the double-stranded DNA and have to be resolved to permit telomere neosynthesis [172, 175]. Unresolved G-quadruplex formations would probably inflict a structural barrier to DNA replication and could be a potential source of genomic instability [20, 175]. G-quadruplexes inhibit telomerase activity and if stabilized by a chemical compound, can act synergistically with Camptothecin or PARP-1 inhibitors to suppress tumor growth in mice [176, 177].

The budding yeast Cell division control protein 13 (Cdc13) and the Repressor-activator protein 1 (Rap1), were between the first telomere associating proteins to be discovered [178]. In *S. cerevisiae*, Cdc13 acts as a single stranded telomeric DNA binding protein that controls telomere elongation by telomerase, whereas Rap1 is responsible for the formation of a multi-protein terminal chromosome cap, termed the "Telosome" [179–181]. The Telosome is composed by the silent information regulator proteins Sir2, Sir3, Sir4 and the telomere-length controllers Rif1 and Rif2 [178, 181–183]. In *S. cerevisiae*, Rap1–Rif1 complexes are negative regulators of telomere length [184–187]. In fission yeast *S. pombe*, Rap1 and Rif1 bind to double-stranded telomeric DNA, interact with the telomere repeat-binding protein Taz1, and regulate telomere length and status of telomeric heterochromatin [20, 188].

Rif1 is highly expressed in mouse embryonic stem (ES) cells and germ cells [189, 190]. The primary function of mammalian Rif1 appears to be involved in DNA-damage [191–193]. However, human RIF1 localizes to dysfunctional telomeres and to telomeric DNA clusters in ALT cells [192], whereas mouse Rif1 maintains telomere length homeostasis of embryonic stem cells (ESCs) by mediating sub-terminal heterochromatin silencing [194].

The protective telomeric structure in mammals is conferred by the assembly of the shelterin complex that caps and shelters chromosome ends from being processed as DNA Double Strand Breaks (DSBs) [171]. Human shelterin is composed of three TTAGGG binding subunits (TRF1, TRF2, and POT1) and three interconnecting molecules (TIN2, TPP1, and RAP1) [171]. Multiple modules of the six shelterin proteins bind to double stranded telomeric repeats via the TRF1 and TRF2 Myb domains, while POT1 associates with single stranded telomeric repeats via its Oligonucleotide/Oligosaccharide-Binding (OB) folds [195, 196] (Fig. 10.3a).

Electron microscopy on purified telomeres of diverse origins, revealed that telomeres do not end as linear DNA molecules, but they form lasso-like structures, that were termed T-loops [197]. It is now well established that the T-loop is formed by the invasion of single-stranded G-overhang into double stranded telomeric repeats, to form a three-stranded DNA displacement loop (D-loop), that renders the 3′ end biochemically inaccessible to DNA repair sensors and nucleases [171, 198] (Fig. 10.1d). T-loops can reach several kb in size, whereas the size of the D-loop is limited by the size of the G-overhang [197, 199].

de Lange [200] proposed that the T-loop structure is the most ancestral element of telomere protection. T-loops preceded protein capping and the emergence of telomerase. The formation of the early T-loops, utilized the complex recombination-dependent, replication (RDR) machinery that pre-existed chromosome linearization as a part of the circular DNA repair system [200, 201].

Human shelterin interacts with several components of the recombinatorial DNA repair machinery such as the MRN complex (Mre11/Rad50/Nbs1) to shape and maintain the T-loop [171, 202, 203]. The DNA helicases RTEL1 (regulator of telomere elongation helicase 1), and the Werner syndrome helicase (WRN) resolve T-loops to enable telomere replication or terminal DNA repair [204–206]. Doksani et al. [207] used stochastic optical reconstruction microscopy (STORM) to show that in mouse chromosomes, TRF2 is the only required shelterin component for biogenesis and/or maintenance of T-loops. Depletion of the other mouse shelterin factors such as TRF1, Rap1, or the POT1 proteins (POT1a and POT1b) did not affect the T-loops.

Several combinations of proteins with similarities to the human shelterin components are found in other species [20]. Mouse shelterin is considered to be composed by TRF1, TRF2, Rap1, and two human POT1 homologues POT1a and POT1b [208–210]. Rap1 is the single human shelterin homologue found at budding yeast telomeres, while Ciliate telomeres contain POT1 and TPP1 [20].

From yeast to humans, the independent or combined depletion of Shelterin or Telosome components from the telomeres, has been related to terminal dysfunction, increased rates of telomeric DNA replication stress, aberrant patterns of telomere sister chromatid exchanges (T-SCEs) and activation of telomeric DNA damage responses [14, 20, 210, 211] (Fig. 10.2c). Mammalian shelterin-free telomeres elicit ATM/ATR (ataxia telangiectasia mutated/ataxia telangiectasia and Rad3 related) DNA damage responses and are processed either by canonical or alternative Non Homologous End Joining (NHEJ) or by homologous recombination (HR) [212].

Telomere dysfunction-induced foci (TIF) are nuclear structures formed by the accumulation of DDR factors, such as γ-H2AX or 53BP1, at telomeres rendered dysfunctional by critical DNA repeat shortening, or by depletion of telomere protective factors [213, 214] (Fig. 10.2a). In the absence of 53BP1, shelterin deprived telomeric repeats immediately become targets for nucleolytic degradation [212].

Shelterin components interact with a plethora of DDR proteins, and nuclear factors implicated in telomere metabolism or the perpetuation of mitotic fidelity (Fig. 10.1d). Several of these molecules such as ATM, ATR, ERCC1/XPF, DNA-PK, BRCA1, BRCA2, PARP-2, TANK1 and TANK2 have been implicated in premature aging syndromes, in hereditary cancer predisposition and in sporadic tumors [18, 215–218] others have DNA binding domains such as HMBOX1 and the COUP nuclear orphan receptors [156].

In addition to chromosome end capping and protection, the shelterin components play important regulatory roles in telomere replenishment and homeostasis [20, 161, 171]. Longer telomeres bind more TRF1 and TRF2 factors and elicit a negative feedback for telomerase activity [219]. Human POT1 and its homologue TEBP from *Tetrahymena*, have been shown to be capable to regulate the formation of terminal G-quadruplexes and to control telomere accessibility by telomerase [220–223] (Fig. 10.3a). Loss of Rap1 induces telomere recombination [224] while TRF2 interacts directly with the DDR machinery through ATM [225], and acts together with Ku70/Ku80 to suppress homologous telomeric crossovers [226].

The Heterochromatic Higher Order of Telomeres and "The Telomere Position Effect"

Beyond the formation of protective, telomere specific nucleoprotein structures, the integrity and functionality of eukaryotic chromosomes depend also on large scale epigenetic modifications that increase the architectural and functional complexity of the linear DNA termini [20, 227]. With the exception of lower eukaryotes, in most species telomeric DNA is organized in unusually spaced, tightly packed nucleosomes [181, 228]. Telomeric nucleosomes display higher mobility compared to the nucleosomes structured on average genomic sequences [229, 230]. The peculiar characteristics of the primary telomeric DNA sequences seem to represent a crucial determinant for chromatin organization both in terms of nucleosomal positioning and spacing [20].

Dipterans like *Anopheles* and *Drosophila melanogaster* have evolved unique biological ways for terminal chromosome organization and capping, as well as for telomere length maintenance [231, 232]. In contrast to most organisms studied,

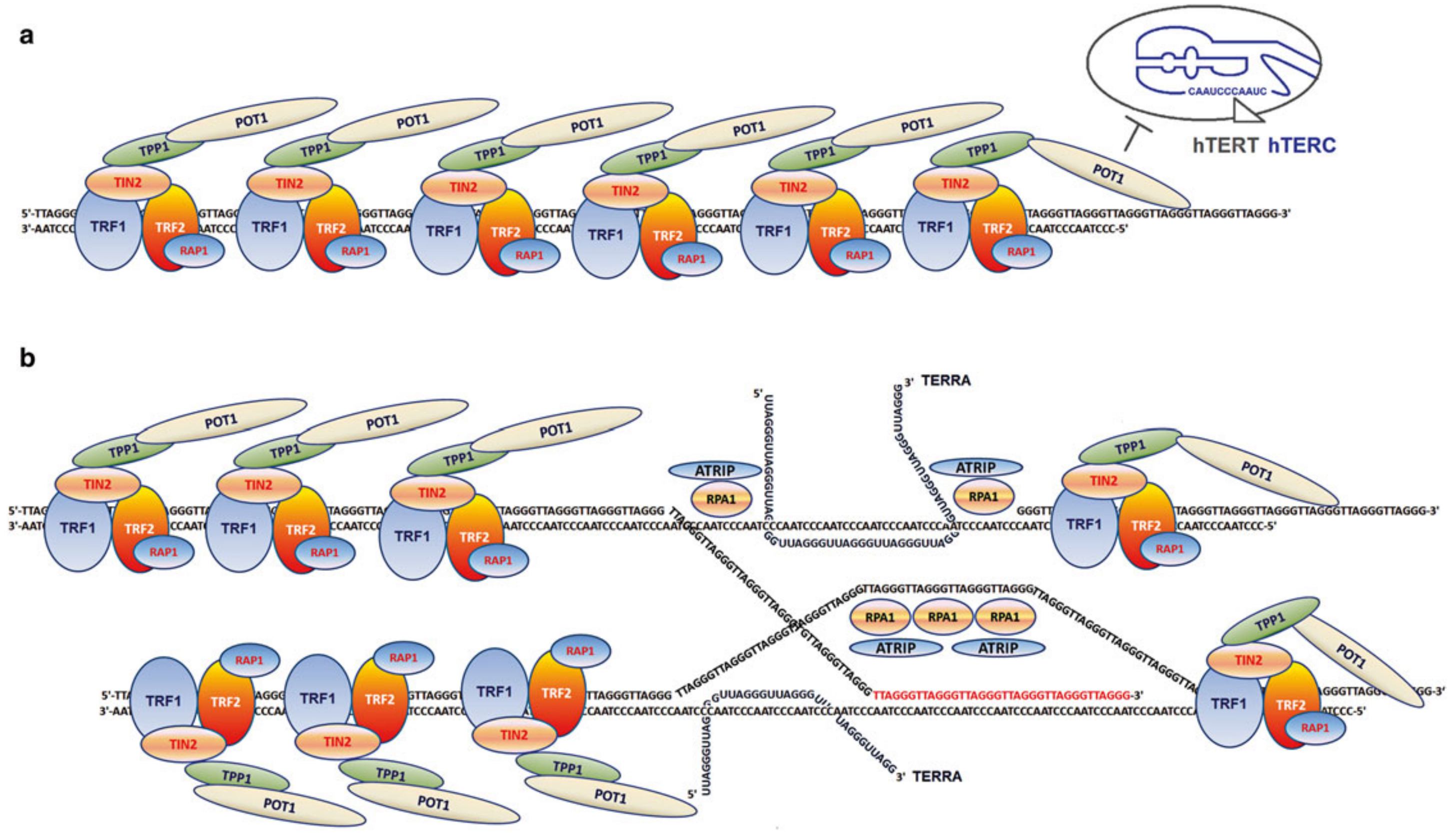


Fig. 10.3 Schematic representation of shelterin binding to human chromosome termini and a hypothetical model for Alternative lengthening of telomeres: Multiple modules of the six known components of human shelterin bind along the whole length of double stranded telomeric repeats via TRF1 and TRF2. POT1 associates with single stranded telomeric DNA of the G-overhang and plays a negative regulatory role in telomere replenishment by telomerase. Longer telomeres bind more TRF1 and TRF2 and elicit a negative feedback for telomerase activity (**a**). A putative Break-Induced-Replication (BIR) model for telomerase-independent telomere elongation: This process generates large segments of single-stranded telomeric DNA that may be protected by single-stranded DNA proteins such as RPA1 and ATRIP and can form DNA–RNA hybrids with TERRAs (*blue*). Neo-synthesized telomeric repeats are depicted in *red color* (**b**)

dipterans do not use telomerase to elongate their telomeres [231, 232]. Telomeres in *Drosophila* are maintained by homologous recombination and gene conversion of the telomere-specific LTR repetitive retro-transposable DNA elements *HeT-A* [231, 233, 234]. Notably the Het-A elements of Drosophila contain sequences that allow the formation of G-quartets in vitro [235]. Fly chromosome termini are capped and protected by the assembly of the "terminin" protein complex [236]. Terminin is composed from the heterochromatin protein 1 (HP1), the HP1/ORC-associated protein HOAP, and the gene products of modigliani (Moi), Ver (Verrocchio), and HipHop [236–240]. The dependence of *Drosophila* chromosome capping on the protein HP1, which is also a major component of mammalian heterochromatin regulation, and the well-established, heterochromatic state of mammalian telomeres suggests a major role of chromatin regulation in the metabolism of the telomeres of linear chromosomes [137, 238].

The heteropyknotic, highly heterochromatic nature of telomeres of several mammalian species was recognized at the early years of chromosome labelling research (1960–1970s) when B. Dutrillaux, T.C. Hsu, J.M. Scheres, J. Lejeune and other pioneer cytogeneticists discovered the C- and T-Banding technologies that were efficiently staining centromeric and telomeric heterochromatin in interphase nuclei and metaphase chromosomes [9, 241–246].

In close proximity to the telomeres, eukaryotic chromosomes display additional structural territories that are termed "sub-telomeres" [247–249]. The sub-terminal DNA regions are located immediately adjacent to telomeric repeats and are comprised of different types of repetitive genomic elements [250]. In humans, subtelomeres contain canonical and degenerate telomeric repeats, and they are highly polymorphic because they frequently undergo large segmental duplications [249].

In contrast to yeast and dipterans, mammalian telomeres and subtelomeric regions, accumulate repressive histone modifications and display extensive hypermethylation of subtelomeric DNA [20, 227, 250, 251]. Mouse subtelomeres are enriched for the H3K9m3 heterochromatin mark, mediated by the Suv39h1 and Suv39h2 histone methyl-transferases [137]. In mice, Rif1 maintains H3K9me3 levels at subtelomeric regions through the negative modulation of the expression of Zscan4 [194]. In humans, DNA methylation accumulates at highly repetitive chromosome territories such as centromeres, pericentric regions and subtelomeres and is considered to act as a suppressor of illegitimate recombination [252–254]. Mammalian telomeres display trimethylated lysines in histones H3 and H4, extensive histone hypoacetylation, accumulation of HP1 and hypermethylation of subtelomeric CpG islands [227, 255]. The heterochromatic state of telomeres is believed to play important roles in the organization of nuclear architecture; it may also contribute to interphase chromatin interactions and the regulation of the mechanisms of telomere replenishment [227, 256].

Telomeres not only constitute targets of epigenetic modifications but they also act as epigenetic agents per se, via a mechanism that is capable to spread heterochromatic silencing to nearby euchromatin [257, 258]. The so called Telomere Position Effect (TPE) regulates the transcriptional activity of genes adjacent to telomere ends, by repressing their expression [259, 260]. TPE is extending in a continuously decreasing fashion from the telomeres to the centromeric chromosomal regions [261]. Telomere position effects have been identified in a variety of lower organisms such as *Schizosaccharomyces pompe*, *Trypanosoma brucei*, *Plasmodium falciparum*, as well as in plants, mice and humans [255, 259, 262–264].

The first organism in which TPE was described was *Drosophila melanogaster* [265–267]. However the TPE phenomenon has been thoroughly investigated and best understood in *Saccharomyces cerevisiae* [266–268]. More than 50 proteins are implicated in *S. Cerevisiae* TPE, with the Ku heterodimer (yKu70p and yKu80p), the Sir-complex (Sir2p, Sir3p and Sir4p) and the C-terminal domain of Rap15 being essential TPE components, whereas in *S. Pompe*, TPE depends on Taz1p, spRap1 (homolog of Rap1p) and Swi6 (ortholog of HP1) [185, 188, 255, 257, 269–271].

Little is known about the molecular mechanism of TPE in Homo sapiens. The histone deacetylase, SIRT6 is considered essential for maintenance of TPE in human cells [272]. Furthermore, the heterochromatin protein HP1, the chromatin remodeling factor SAL1 and the shelterin components TRF1 and TIN2 are thought to be key players of the human TPE processes [273, 274]. In humans, TPE was first systematically studied in HeLa cells, in which an exogenous reporter gene was stochastically incorporated into the genome. Clones with the exogenous gene inserted adjacent to telomeres, presented tenfold decreased expression, in contrast to clones carrying the reporter in random genomic positions, with the phenomenon being telomere length and heterochromatin formation dependent. In this context, hTERT overexpression that results in telomere elongation, led to further decrease of the adjacent transgene's expression, while overexpression of TRF1 or treatment with the histone deacetylation inhibitor Trichostatin A, restored expression [259, 275].

The biological machinery that regulates the expression of subtelomeric genes may be implicated in normal human ageing and nosology and especially in age related diseases when telomeres are substantially shorter [259]. A rare myopathy termed Facio-Scapulo-Humeral Dystrophy (FSHD) is currently the only human disorder directly associated to TPE. The leading causative factor of FSHD, is the DUX4 homeobox protein, expressed by a gene located adjacent to the subtelomeric D4Z4 tandem repeat array, within chromosome band 4q35 [276]. Normal FSHD alleles carry 11–110 copies of the D4Z4 repeat, that are acting as a barrier of DUX4 expression, whereas pathogenic alleles have only 1–10 tandem repeats and allow the expression of DUX4 [258, 276–281]. Stadler et al. [258] have shown that the expression of DUX4 can be further up-regulated in FSHD myoblasts and myotubes with short telomeres.

TERRAs: Novel Partners of Telomere Homeostasis

In 2007, Azzalin et al. [282] brought down one longstanding dogma of molecular biology: In contrast to the "good-chromatin" euchromatin, that is "unpacked" and capable to be transcribed into coding or non-coding RNAs, highly heterochromatic regions such as the mammalian centromeres and subtelomeres, were generally considered transcriptionally silent. The groups of J. Lingner in 2007, and M. Blasco (2008) showed that throughout eukaryotes, subtelomeric regions and telomeres are transcribed into variable sized, long non-coding telomeric repeat-containing RNAs that were termed TERRAs [282–285]. From anthozoans to humans, TERRAs consist of subtelomeric sequences and multiple tracts of the hexa-nucleotide repeat (5′-UUAGGG-3′) [283]. TERRA molecules are considered to be much shorter than their available C-rich telomeric DNA template and were shown to display heterogeneous lengths between species and chromosomes, ranging from 100 bp, up to more than 9 kb [178, 282, 285–288].

It is very possible that the eukaryote telomeric transcriptome will be expanded with several types of telomeric non coding RNA transcripts: In addition to TERRA, *Schizosaccharomyces pombe* telomeres express several types of telomeric transcripts named ARIA, ARRET and anti-ARRET [289, 290]. ARIAs were also found in plants [291]. They are mainly comprised by C-rich complementary RNA sequences that use the G-rich telomeric DNA as template, suggesting that in parallel to subtelomeric regions, canonical telomeric DNA repeats can serve as transcription start-sites for RNA polymerases [289, 290]. ARRET display sequence complementarity with proximal subtelomeric regions and lack canonical telomere repeats [289]. The anti-ARRETs are transcribed by the antiparallel strand of the ARRET template hence they are complementary to ARRET [283, 289].

Up to date, TERRA is the most well studied species of the telomere transcriptome [283]. Chromatin immunoprecipitation (ChIP) experiments revealed that from fungi to humans, TERRAs are neosynthesized by the DNA-dependent RNA polymerase II (RNAPII) that produces G-rich, TERRA molecules, using the C-rich telomeric DNA strand as a template [282, 285]. A proportion of TERRAs is modified at their 3′-end, by the addition of a polyadenylation tag [292, 293]. In human cells polyadenylated TERRAs fractionate within the nucleoplasmic fraction and do not associate with chromatin [293]. It is estimated that around 7 % of human TERRA is polyadenylated [292]. The larger fraction of human TERRAs is non-polyadenylated and found associated with telomeric heterochromatin [293].

Human CpG dinucleotide-containing TERRA promoters were found in at least half of the highly heterochromatic human subtelomeres [282]. TERRA promoters are active during G1 and G2 phases of the cell cycle and silenced during the S-phase [293]. Subtelomeric DNA methylation and other subterminal epigenetic modifications are direct modulators of TERRA metabolism. Combined depletion of the DNA methyltransferases DNMT1 and DNMT3b that control heterochromatin state of the subterminal CpG islands of human chromosomes, led to up-regulation of TERRA transcription in diploid and cancer cell lines [294]. The shelterin component TRF2 is considered a negative controller of TERRAs. Telomere de-protection induced by TRF2 knockdown, leads to overproduction of TERRA in human fibroblasts and HeLa cells [283, 295, 296]. On the contrary, Deng et al. [297] have shown that the chromatin organizing factor CTCF and the cohesin subunit Rad21 bind to subtelomeric regions and promote TERRA transcription in human cell lines. In budding yeast, Rif1 and Rif2 block TERRA transcription at all telomeres whereas the proteins of the Sir-family are involved in TERRA repression only at the subset of telomeres carrying the conserved yeast subtelomere structures termed Y'-elements [283, 298].

The formation of DNA/RNA hybrids named R-loops is a rare natural outcome of transcription, caused by invasion of the DNA double strand, by nascent RNA transcripts. R-loops have been considered causative factors of genome fragility and can induce repressive chromatin marks over mammalian gene terminators [299]. Co-transcriptional base-complementarity may promote the formation of DNA/RNA hybrids at telomeres [300, 301]. G-rich telomeric DNA/RNA hybrids are capable to form G-quadruplexes and thus they can inhibit telomerase accessibility to telomeres [283].

In yeast, the poly(A)-polymerase Pap1p polyadenylates and stabilizes TERRAs whereas overexpression of the 5′-3′ exonuclease Rat1p, is capable to fully degrade TERRAs from the nucleus [302]. Yeast defective for Rat1p, accumulate TERRAs and display some degree of telomere shortening, suggesting that telomere-associated TERRAs may directly affect the mechanisms of telomere replenishment [283]. Telomere attrition in Rat1 deficient cells was rescued by overexpression of RNaseH, indicating that telomere metabolism is affected by the formation of telomeric RNA/DNA hybrids [284]. In human cell lines utilizing the alternative lengthening of telomeres (ALT-pathway) depletion of RNaseH1 caused TERRA-telomeric hybrid accumulation, exposure of single-stranded telomeric DNA, increased levels of the single strand DNA binding protein RPA and abrupt telomere length excision, whereas overexpression of RNaseH1 suppressed ALT telomeric recombination and led to telomere shortening [300]. Analogous changes at the levels of RNaseH1 in telomerase positive cells did not demonstrate any of the above phenotypes suggesting a specific role for TERRA in alternative lengthening of telomeres [300].

TERRA association with human telomeres is regulated by proteins known to participate in the biological processes of mRNA decay (NMD), such as the RNA/DNA helicase and ATPase UPF1, the RNA endonuclease hEST1A/SMG6, and the protein kinase SMG1 [282, 283]. Depletion of UPF1 in telomerase positive HeLa cells affected replication of the leading telomere strand causing telomeric fragility [303]. UPF1-depleted cells accumulate telomeric R-loops because the C-rich telomeric strand acts as the template for both TERRA and leading-strand DNA replication [303].

Telomeropathies

The first premature aging disorder linked to impaired telomere metabolism was Dyskeratosis Congenita (DC), an inherited bone marrow failure condition, first described in 1906 by Zinsser [304]. DC is a rare, childhood onset disease, estimated to occur with a frequency of 1, in one million individuals, with death occurring at an average age of 16 [305]. The three distinctive clinical characteristics of DC are skin pigmentation abnormalities, nail dystrophy and mucosal leukoplakia [306, 307]. These phenotypes are present in almost 80–90 % of diagnosed cases [307].

Organ failure is a typical outcome of DC patients, with bone marrow in 80 % of the cases being the first tissue to be affected, leading to aplastic anemia [306, 308]. In addition to bone marrow failure, pulmonary fibrosis and cancer, are the main fatal complications of the disease [306]. The most frequent malignancies are myelodysplastic syndromes (MDS), head and neck cancers (squamous carcinomas), and acute myeloid leukemias (AML) [115, 309]. The clinical phenotype of DC is expanding continuously with a plethora of disease manifestations such as immune deficiency, cardiomyopathy, liver cirrhosis, mental retardation, epiphora (excessive tears in the eyes), dental loss, osteoporosis, premature hair loss/greying and deafness [15, 19].

DC patients, present abnormally short telomeres as compared to individuals belonging in the same chronological age group [114, 310]. The premature telomere attrition is considered to be the underlying cause behind most of the pathological features of DC signifying this disease as a "Telomeropathy" [311–314]. This is supported by the fact that all major mutations linked to DC, involved genes implicated in telomere maintenance: About 50 % of DC patients have a mutation in one of the three main components of the telomerase holoenzyme complex, Dyskerin, TERC and TERT [315]. The DKC1 gene, that encodes Dyskerin protein, is located at the X chromosome [316]. Mutations of DKC1 cause a recessive, X-linked form of DC characterized by a severe phenotype [316]. An autosomal dominant milder form of DC, results from defects of the gene responsible for TERC, whereas mutations in the reverse transcriptase component TERT, have been linked to autosomal dominant and autosomal recessive inheritance [311, 314, 317, 318]. Many TERT mutations affect severely the activity of telomerase [311, 319–321]. However, in some cases the phenotype appears relatively milder therefore they are thought to behave more as risk polymorphisms, than as determinants of disease [312] (Table 10.1).

Homozygous mutations of the NOP10 and NHP2 components of telomerase holoenzyme are considered responsible for an autosomal recessive form of DC [322, 323]. Approximately 11 % of DC cases result from mutations in the TINF2 gene, encoding the Shelterin component TIN2 [171, 324]. TIN-2 telomeropathy is linked to a severe clinical phenotype and transmitted in an autosomal dominant mode [325]. While five out of six DC genes result in impaired telomerase activity, TIN2 mutations are thought to jeopardize either the protection of telomeres or the accessibility of telomere ends by telomerase, leading to more pronounced telomere attrition [312]. Most TINF2 mutation patients, display exceptionally short telomeres from a much earlier age than the patients bearing any other mutant DC gene [312].

Dyskeratosis congenita is considered a highly heterogeneous disorder with its genetic cause and phenotype overlapping significantly with various syndromes such as Hoyeraal-Hreidersson (HHS), Coats-Plus, and Revesz syndrome. HHS particularly, is believed to be a severe DC form and the phenotype diverges by the addition of cerebellar hypoplasia, microcephaly and intrauterine growth retardation [326]. Similarly to HHS, Revesz syndrome is also extremely rare and is presenting with HHS symptoms and bilateral exudative retinopathy [327]. Patients of those severe telomeropathies exert extremely short telomeres and a higher mortality rate than

Table 10.1 Telomeropoathies and implicated genes

Genes involved	Process affected	Role in telomere metabolism	Diseases	References
TERC	Telomerase activity	Telomere length maintenance	AD-DC, Sporadic & Familial IPF, AA, MDS	[315, 318, 320, 321, 336–338, 344–346]
TERT	Telomerase activity	Telomere length maintenance	AD-DC, Sporadic & Familial IPF, AA, MDS	[311, 314, 315, 317, 320, 321, 336–338, 344–346]
DKC1	Telomerase activity	Telomere length maintenance	XR DC, HHS	[315, 316]
CTC1	Telomere capping/CST complex	Telomere protection	Coats-Plus syndrome	[328, 329]
TINF2	Shelterin	Telomere regulation	AD-DC, HHS, Revesz syndrome	[171, 312, 324, 325]
NOP10	Telomerase activity	Telomere length maintenance	AR-DC	[323]
NHP2	Telomerase activity	Telomere length maintenance	AR-DC	[322]
TCAB1	Telomerase activity	Telomerase assembly	AR-DC	[350]
FANCD2	Telomere maintenance	Telomere regulation	FA	[347, 348]
RECQL4	D-loop resolution	Telomere regulation	RT	[349]
NBN	MRN complex	Telomere length maintenance	NBS	[53, 148, 350]
ATM	Telomeric DDR?	Telomere regulation	AT	[351–353]
BLM	DNA unwinding?	Telomere regulation	BS	[359]
WRN	D-loop resolution	Telomere regulation	WS	[358]
DNMT3B	Heterochromatinization	Subtelomere/Telomere regulation	ICF	[360–364]

AD/AR/XR-DC autosomal dominant/autosomal recessive/X-linked recessive, dyskeratosis congenita, *IPF* idiopathic pulmonary fibrosis, *AA* aplastic Anemia, *MDS* myelodysplastic syndromes, *HHS* Hoyeraal-Hreidersson syndrome, *FA* Fanconi Anemia, *RT* Rothmund Thomson, *NBS* Nijmegen Breakage syndrome, *AT* ataxia telangiectasia, *BS* bloom syndrome, *WS* Werner syndrome, *ICF* immunodeficiency centromere instability and facial anomalies syndrome, *DDR* DNA damage response classic DC cases. Coats-Plus syndrome has been linked to CTC1 mutations and patients exhibit exudative retinopathy and intracranial calcifications, features that are also observed in HHS and Revesz syndromes [328, 329]. Due to the small number of HHS, Revesz and Coats-Plus syndrome patients and their overlapping clinical findings, it is extremely difficult to determine guidelines distinguishing the three disorders [330].

The most common adulthood onset manifestation of impaired telomere maintenance is the idiopathic pulmonary fibrosis (IPF) which is a fatal progressive lung disease that typically presents in the fifth decade with the incidence increasing with advanced age [50, 331–333]. IPF is characterized as sporadic and only 0.5–2.2 % of the cases present familial inheritance [334, 335]. Mutations in TERT and TERC telomerase subunits have been identified in about 1–3 % of sporadic and 8–15 % of familial IPF cases [321, 336–338]. The pattern of inheritance in most families is autosomal dominant and consistent with incomplete penetrance and haploinsufficiency of telomerase [335, 339–342]. Furthermore, 37 % of familial and 25 % of sporadic cases present telomere length shorter than the 1/10 of the general population, suggesting an association with impaired telomere replenishment. This phenomenon may be explained by the underlying telomerase mutations and/or the effects of environmental causes such as smoking, which is a common characteristic of at least 50 % of IPF patients [342]. Because telomerase is expressed only in mitotically active cells, IPF may result, partly, from the senescence or loss of a stem cell population in lungs, capable to respond to continual injuries over time. Due to the haploinsufficiency of telomerase, the clinical phenotype arises only after adequate time has elapsed and thus the cells have conducted enough division rounds to present critically short telomeres [321, 343]. Mutations in TERT and TERC have also been found in bone morrow failure disorders other than DC, such as the myelodysplastic syndromes (MDS) and aplastic anemia (AA) [320, 344–346]. Another adult-onset disorder presenting telomerase mutations is familial liver cirrhosis, which is a known complication of dyskeratosis congenita [116, 311].

Dysfunctional or prematurely shortened telomeres have been observed in several premature ageing or cancer pre-disposing syndromes caused by mutations in genes encoding proteins that are key players of DNA damage responses and are also implicated in telomere metabolism [347]. Telomere dysfunction foci (TIFs) and short telomeres have been found in lymphocytes from patients with Fanconi anemia (FA) [348]. A subset of FA patients, carry mutations in FANCD2, a protein that interacts with telomeric DNA and regulates the levels of the shelterin component TRF1 [347]. Mutations in the RECQL4 helicase that resolves telomeric D-loops, are the causal factors of the Rothmund–Thomson progeroid syndrome (RT), a disease with some DC clinical characteristics that is accompanied by increased TIF frequencies and excessive telomeric instability [349]. The Nijmegen chromosome breakage syndrome occurs due to mutations of the Nibrin (NBN) gene that encodes a member of the MRN (MRE11/RAD50/NBN) complex [53]. The MRN multimer is involved in the mechanisms regulating telomere length maintenance [148, 350]. Cells from patients with Nijmegen syndrome show decreased telomere length and poor proliferation rates [53].

Ataxia telangiectasia (AT) is characterized by neurological deterioration, immunodeficiency, spontaneous chromosomal instability, hypersensitivity to ionizing radiation, predisposition to leukemias and lymphomas and premature ageing [351]. In yeast, the Ataxia telangiectasia mutated (Atm) homolog, Tel1, is necessary for normal telomere length regulation [352]. Lymphocytes of AT patients exert accelerated telomere shortening and frequent terminal chromosome fusions [351, 353].

Two other rare genetic cancer predisposition diseases Werner's syndrome (WS), and Bloom's syndrome (BS), are related to premature aging [354]. WS and BS are caused by loss of function of the RecQ helicases WRN and BLM, respectively [355, 356]. Both diseases are characterized by replication errors, hyper-DNA recombination and chromosomal instability [355, 357]. Mutations in the WRN gene were associated with insufficient replication of the G-rich telomeric strand [358]. BLM also contributes to chromosome-end maintenance through its genome-wide activity in resolving difficult-to-replicate regions such as telomeres. BLM-deficient normal human fibroblast cells display elevated frequencies of telomere dysfunction [359].

The Immunodeficiency, Centromeric region instability and Facial anomalies syndrome (ICF) is a rare condition caused by mutations in the DNA methylotransferase gene DNMT3B [360]. Cells from ICF patient's display telomeric abnormalities and reduced subtelomeric methylation [361–364]. Remarkably, ICF cells are characterized by abnormally high levels of TERRA transcripts suggesting the first link between alterations in the telomere transcriptome and human nosology [365].

References

1. Jackson SP, Bartek J (2009) The DNA-damage response in human biology and disease. Nature 461(7267):1071–1078
2. Jackson SP (2009) The DNA-damage response: new molecular insights and new approaches to cancer therapy. Biochem Soc Trans 37(Pt 3):483–494
3. Ghosal G, Chen J (2013) DNA damage tolerance: a double-edged sword guarding the genome. Transl Cancer Res 2(3):107–129
4. Fry RC, Begley TJ, Samson LD (2005) Genome-wide responses to DNA-damaging agents. Annu Rev Microbiol 59:357–377
5. Kreuzer KN (2013) DNA damage responses in prokaryotes: regulating gene expression, modulating growth patterns, and manipulating replication forks. Cold Spring Harb Perspect Biol 5(11):a012674
6. Ciccia A, Elledge SJ (2010) The DNA damage response: making it safe to play with knives. Mol Cell 40(2):179–204
7. Stracker TH, Roig I, Knobel PA, Marjanovic M (2013) The ATM signaling network in development and disease. Front Genet 4:37
8. Jackson SP, Durocher D (2013) Regulation of DNA damage responses by ubiquitin and SUMO. Mol Cell 49(5):795–807
9. Goodpasture C, Bloom SE, Hsu TC, Arrighi FE (1976) Human nucleolus organizers: the satellites or the stalks? Am J Hum Genet 28(6):559–566
10. Garavis M, Gonzalez C, Villasante A (2013) On the origin of the eukaryotic chromosome: the role of noncanonical DNA structures in telomere evolution. Genome Biol Evol 5(6): 1142–1150

11. Olovnikov AM (1971) Principle of marginotomy in template synthesis of polynucleotides. Dokl Akad Nauk SSSR 201(6):1496–1499
12. Olovnikov AM (1973) A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. J Theor Biol 41(1):181–190
13. Watson JD (1972) Origin of concatemeric T7 DNA. Nat New Biol 239(94):197–201
14. De Lange T (2005) Telomere-related genome instability in cancer. Cold Spring Harb Symp Quant Biol 70:197–204
15. O'Sullivan RJ, Karlseder J (2010) Telomeres: protecting chromosomes against genome instability. Nat Rev Mol Cell Biol 11(3):171–181
16. Egan ED, Collins K (2012) Biogenesis of telomerase ribonucleoproteins. RNA 18(10): 1747–1759
17. Levy MZ, Allsopp RC, Futcher AB, Greider CW, Harley CB (1992) Telomere end-replication problem and cell aging. J Mol Biol 225(4):951–960
18. Palm W, de Lange T (2008) How shelterin protects mammalian telomeres. Annu Rev Genet 42:301–334
19. de Lange T (2009) How telomeres solve the end-protection problem. Science 326(5955): 948–952
20. Giraud-Panis MJ, Pisano S, Poulet A, Le Du MH, Gilson E (2010) Structural identity of telomeric complexes. FEBS Lett 584(17):3785–3799
21. Muller HJ (1938) The remaking of chromosomes. Collecting Net 13:181–198
22. McClintock B (1939) The behavior in successive nuclear divisions of a chromosome broken at meiosis. Proc Natl Acad Sci U S A 25(8):405–416
23. McClintock B (1941) The stability of broken ends of chromosomes in Zea Mays. Genetics 26(2):234–282
24. McClintock B (1938) The production of homozygous deficient tissues with mutant characteristics by means of the aberrant mitotic behavior of ring-shaped chromosomes. Genetics 23(4):315–376
25. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. Nature 396(6712):643–649
26. Gisselsson D, Pettersson L, Hoglund M, Heidenblad M, Gorunova L, Wiegant J et al (2000) Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. Proc Natl Acad Sci U S A 97(10):5357–5362
27. Gagos S, Irminger-Finger I (2005) Chromosome instability in neoplasia: chaotic roots to continuous growth. Int J Biochem Cell Biol 37(5):1014–1033
28. Murnane JP (2012) Telomere dysfunction and chromosome instability. Mutat Res 730(1-2):28–36
29. Gisselsson D, Hoglund M (2005) Connecting mitotic instability and chromosome aberrations in cancer--can telomeres bridge the gap? Semin Cancer Biol 15(1):13–23
30. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674
31. Blackburn EH, Gall JG (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. J Mol Biol 120(1):33–53
32. Szostak JW, Blackburn EH (1982) Cloning yeast telomeres on linear plasmid vectors. Cell 29(1):245–255
33. Greider CW, Blackburn EH (1987) The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell 51(6):887–898
34. Greider CW, Blackburn EH (1989) A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. Nature 337(6205):331–337
35. Morin GB (1989) The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. Cell 59(3):521–529
36. Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD et al (1988) A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. Proc Natl Acad Sci U S A 85(18):6622–6626

37. Lorite P, Carrillo JA, Palomeque T (2002) Conservation of (TTAGG)(n) telomeric sequences among ants (Hymenoptera, Formicidae). J Hered 93(4):282–285
38. Zielke S, Bodnar A (2010) Telomeres and telomerase activity in scleractinian corals and Symbiodinium spp. Biol Bull 218(2):113–121
39. Gomes NM, Ryder OA, Houck ML, Charter SJ, Walker W, Forsyth NR et al (2011) Comparative biology of mammalian telomeres: hypotheses on ancestral states and the roles of telomeres in longevity determination. Aging Cell 10(5):761–768
40. Zakian VA (1995) Telomeres: beginning to understand the end. Science 270(5242):1601–1607
41. Makarov VL, Hirose Y, Langmore JP (1997) Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. Cell 88(5):657–666
42. Chai W, Shay JW, Wright WE (2005) Human telomeres maintain their overhang length at senescence. Mol Cell Biol 25(6):2158–2168
43. Bischoff C, Graakjaer J, Petersen HC, Hjelmborg J, Vaupel JW, Bohr V et al (2005) The heritability of telomere length among the elderly and oldest-old. Twin Res Hum Genet 8(5):433–439
44. Forsyth NR, Elder FF, Shay JW, Wright WE (2005) Lagomorphs (rabbits, pikas and hares) do not use telomere-directed replicative aging in vitro. Mech Ageing Dev 126(6-7):685–691
45. Raices M, Verdun RE, Compton SA, Haggblom CI, Griffith JD, Dillin A et al (2008) C. elegans telomeres contain G-strand and C-strand overhangs that are bound by distinct proteins. Cell 132(5):745–757
46. Oganesian L, Karlseder J (2011) Mammalian 5′ C-rich telomeric overhangs are a mark of recombination-dependent telomere maintenance. Mol Cell 42(2):224–236
47. Hayflick L (1997) Mortality and immortality at the cellular level. A review. Biochemistry (Mosc) 62(11):1180–1190
48. Hayflick L, Moorhead PS (1961) The serial cultivation of human diploid cell strains. Exp Cell Res 25:585–621
49. Harley CB, Futcher AB, Greider CW (1990) Telomeres shorten during ageing of human fibroblasts. Nature 345(6274):458–460
50. Armanios M, Blackburn EH (2012) The telomere syndromes. Nat Rev Genet 13(10):693–704
51. Donate LE, Blasco MA (2011) Telomeres in cancer and ageing. Philos Trans R Soc Lond B Biol Sci 366(1561):76–84
52. Ahmed A, Tollefsbol T (2001) Telomeres and telomerase: basic science implications for aging. J Am Geriatr Soc 49(8):1105–1109
53. Ranganathan V, Heine WF, Ciccone DN, Rudolph KL, Wu X, Chang S et al (2001) Rescue of a telomere length defect of Nijmegen breakage syndrome cells requires NBS and telomerase catalytic subunit. Curr Biol 11(12):962–966
54. Benetos A, Okuda K, Lajemi M, Kimura M, Thomas F, Skurnick J et al (2001) Telomere length as an indicator of biological aging: the gender effect and relation with pulse pressure and pulse wave velocity. Hypertension 37(2 Pt 2):381–385
55. Ding Z, Mangino M, Aviv A, Spector T, Durbin R (2014) Estimating telomere length from whole genome sequence data. Nucleic Acids Res 42(9):e75
56. Hastie ND, Dempster M, Dunlop MG, Thompson AM, Green DK, Allshire RC (1990) Telomere reduction in human colorectal carcinoma and with ageing. Nature 346(6287):866–868
57. Kimura M, Stone RC, Hunt SC, Skurnick J, Lu X, Cao X et al (2010) Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. Nat Protoc 5(9):1596–1607
58. Halazonetis TD, Gorgoulis VG, Bartek J (2008) An oncogene-induced DNA damage model for cancer development. Science 319(5868):1352–1355
59. Shay JW, Wright WE (2011) Role of telomeres and telomerase in cancer. Semin Cancer Biol 21(6):349–353

60. Greider CW (1996) Telomere length regulation. Annu Rev Biochem 65:337–365
61. Sherr CJ, DePinho RA (2000) Cellular senescence: mitotic clock or culture shock? Cell 102(4):407–410
62. Mathon NF, Lloyd AC (2001) Cell senescence and cancer. Nat Rev Cancer 1(3):203–213
63. Rodier F, Campisi J (2011) Four faces of cellular senescence. J Cell Biol 192(4):547–556
64. Shay JW, Roninson IB (2004) Hallmarks of senescence in carcinogenesis and cancer therapy. Oncogene 23(16):2919–2933
65. Falandry C, Bonnefoy M, Freyer G, Gilson E (2014) Biology of cancer and aging: a complex association with cellular senescence. J Clin Oncol 32(24):2604–2610
66. Bodnar AG, Ouellette M, Frolkis M, Holt SE, Chiu CP, Morin GB et al (1998) Extension of life-span by introduction of telomerase into normal human cells. Science 279(5349): 349–352
67. Harrington L (2012) Haploinsufficiency and telomere length homeostasis. Mutat Res 730(1-2):37–42
68. Weinrich SL, Pruzan R, Ma L, Ouellette M, Tesmer VM, Holt SE et al (1997) Reconstitution of human telomerase with the template RNA component hTR and the catalytic protein subunit hTRT. Nat Genet 17(4):498–502
69. Harrington L, McPhail T, Mar V, Zhou W, Oulton R, Bass MB et al (1997) A mammalian telomerase-associated protein. Science 275(5302):973–977
70. Harrington L, Zhou W, McPhail T, Oulton R, Yeung DS, Mar V et al (1997) Human telomerase contains evolutionarily conserved catalytic and structural subunits. Genes Dev 11(23):3109–3115
71. Hughes TR, Evans SK, Weilbaecher RG, Lundblad V (2000) The Est3 protein is a subunit of yeast telomerase. Curr Biol 10(13):809–812
72. Lin JJ, Zakian VA (1995) An in vitro assay for Saccharomyces telomerase requires EST1. Cell 81(7):1127–1135
73. Lingner J, Cech TR, Hughes TR, Lundblad V (1997) Three Ever Shorter Telomere (EST) genes are dispensable for in vitro yeast telomerase activity. Proc Natl Acad Sci U S A 94(21):11190–11195
74. Lundblad V, Blackburn EH (1993) An alternative pathway for yeast telomere maintenance rescues est1- senescence. Cell 73(2):347–360
75. Lundblad V, Szostak JW (1989) A mutant with a defect in telomere elongation leads to senescence in yeast. Cell 57(4):633–643
76. Snow BE, Erdmann N, Cruickshank J, Goldman H, Gill RM, Robinson MO et al (2003) Functional conservation of the telomerase protein Est1p in humans. Curr Biol 13(8): 698–704
77. Reichenbach P, Hoss M, Azzalin CM, Nabholz M, Bucher P, Lingner J (2003) A human homolog of yeast Est1 associates with telomerase and uncaps chromosome ends when overexpressed. Curr Biol 13(7):568–574
78. Fukuhara N, Ebert J, Unterholzner L, Lindner D, Izaurralde E, Conti E (2005) SMG7 is a 14-3-3-like adaptor in the nonsense-mediated mRNA decay pathway. Mol Cell 17(4): 537–547
79. Noel JF, Larose S, Abou Elela S, Wellinger RJ (2012) Budding yeast telomerase RNA transcription termination is dictated by the Nrd1/Nab3 non-coding RNA termination pathway. Nucleic Acids Res 40(12):5625–5636
80. Sauerwald A, Sandin S, Cristofari G, Scheres SH, Lingner J, Rhodes D (2013) Structure of active dimeric human telomerase. Nat Struct Mol Biol 20(4):454–460
81. Tuzon CT, Wu Y, Chan A, Zakian VA (2011) The Saccharomyces cerevisiae telomerase subunit Est3 binds telomeres in a cell cycle- and Est1-dependent manner and interacts directly with Est1 in vitro. PLoS Genet 7(5):e1002060
82. Witkin KL, Collins K (2004) Holoenzyme proteins required for the physiological assembly and activity of telomerase. Genes Dev 18(10):1107–1118
83. Min B, Collins K (2009) An RPA-related sequence-specific DNA-binding subunit of telomerase holoenzyme is required for elongation processivity and telomere maintenance. Mol Cell 36(4):609–619

84. Teixeira MT, Arneric M, Sperisen P, Lingner J (2004) Telomere length homeostasis is achieved via a switch between telomerase- extendible and -nonextendible states. Cell 117(3):323–335
85. Chen JL, Greider CW (2004) An emerging consensus for telomerase RNA structure. Proc Natl Acad Sci U S A 101(41):14683–14684
86. Cristofari G, Lingner J (2003) Fingering the ends: how to make new telomeres. Cell 113(5):552–554
87. Feng J, Funk WD, Wang SS, Weinrich SL, Avilion AA, Chiu CP et al (1995) The RNA component of human telomerase. Science 269(5228):1236–1241
88. Fu D, Collins K (2007) Purification of human telomerase complexes identifies factors involved in telomerase biogenesis and telomere length regulation. Mol Cell 28(5):773–785
89. Grozdanov PN, Roy S, Kittur N, Meier UT (2009) SHQ1 is required prior to NAF1 for assembly of H/ACA small nucleolar and telomerase RNPs. RNA 15(6):1188–1197
90. Kanemaki M, Kurokawa Y, Matsu-ura T, Makino Y, Masani A, Okazaki K et al (1999) TIP49b, a new RuvB-like DNA helicase, is included in a complex together with another RuvB-like DNA helicase, TIP49a. J Biol Chem 274(32):22437–22444
91. Makino Y, Kanemaki M, Kurokawa Y, Koji T, Tamura T (1999) A rat RuvB-like protein, TIP49a, is a germ cell-enriched novel DNA helicase. J Biol Chem 274(22):15329–15335
92. Venteicher AS, Meng Z, Mason PJ, Veenstra TD, Artandi SE (2008) Identification of ATPases pontin and reptin as telomerase components essential for holoenzyme assembly. Cell 132(6):945–957
93. Darzacq X, Kittur N, Roy S, Shav-Tal Y, Singer RH, Meier UT (2006) Stepwise RNP assembly at the site of H/ACA RNA transcription in human cells. J Cell Biol 173(2):207–218
94. Boulon S, Verheggen C, Jady BE, Girard C, Pescia C, Paul C et al (2004) PHAX and CRM1 are required sequentially to transport U3 snoRNA to nucleoli. Mol Cell 16(5):777–787
95. Hukezalie KR, Wong JM (2013) Structure-function relationship and biogenesis regulation of the human telomerase holoenzyme. FEBS J 280(14):3194–3204
96. Jady BE, Bertrand E, Kiss T (2004) Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body-specific localization signal. J Cell Biol 164(5):647–652
97. Tycowski KT, Shu MD, Kukoyi A, Steitz JA (2009) A conserved WD40 protein binds the Cajal body localization signal of scaRNP particles. Mol Cell 34(1):47–57
98. Venteicher AS, Artandi SE (2009) TCAB1: driving telomerase to Cajal bodies. Cell Cycle 8(9):1329–1331
99. Stellwagen AE, Haimberger ZW, Veatch JR, Gottschling DE (2003) Ku interacts with telomerase RNA to promote telomere addition at native and broken chromosome ends. Genes Dev 17(19):2384–2395
100. Sealey DC, Zheng L, Taboski MA, Cruickshank J, Ikura M, Harrington LA (2010) The N-terminus of hTERT contains a DNA-binding domain and is required for telomerase activity and cellular immortalization. Nucleic Acids Res 38(6):2019–2035
101. Greider CW (1991) Telomerase is processive. Mol Cell Biol 11(9):4572–4580
102. Chen JL, Greider CW (2003) Template boundary definition in mammalian telomerase. Genes Dev 17(22):2747–2752
103. Lai CK, Miller MC, Collins K (2003) Roles for RNA in telomerase nucleotide and repeat addition processivity. Mol Cell 11(6):1673–1683
104. Cohn M, Blackburn EH (1995) Telomerase in yeast. Science 269(5222):396–400
105. Prescott J, Blackburn EH (1997) Telomerase RNA mutations in Saccharomyces cerevisiae alter telomerase action and reveal nonprocessivity in vivo and in vitro. Genes Dev 11(4):528–540
106. Prowse KR, Avilion AA, Greider CW (1993) Identification of a nonprocessive telomerase activity from mouse cells. Proc Natl Acad Sci U S A 90(4):1493–1497
107. Pascolo E, Wenz C, Lingner J, Hauel N, Priepke H, Kauffmann I et al (2002) Mechanism of human telomerase inhibition by BIBR1532, a synthetic, non-nucleosidic drug candidate. J Biol Chem 277(18):15566–15572

108. Skvortsov DA, Zvereva ME, Shpanchenko OV, Dontsova OA (2011) Assays for detection of telomerase activity. Acta Naturae 3(1):48–68
109. Allshire RC, Dempster M, Hastie ND (1989) Human telomeres contain at least three types of G-rich repeat distributed non-randomly. Nucleic Acids Res 17(12):4611–4627
110. Aubert G, Hills M, Lansdorp PM (2012) Telomere length measurement-caveats and a critical assessment of the available technologies and tools. Mutat Res 730(1-2):59–67
111. Wright WE, Brasiskyte D, Piatyszek MA, Shay JW (1996) Experimental elongation of telomeres extends the lifespan of immortal x normal cell hybrids. EMBO J 15(7):1734–1741
112. Aviv A, Hunt SC, Lin J, Cao X, Kimura M, Blackburn E (2011) Impartial comparative analysis of measurement of leukocyte telomere length/DNA content by Southern blots and qPCR. Nucleic Acids Res 39(20):e134
113. Montpetit AJ, Alhareeri AA, Montpetit M, Starkweather AR, Elmore LW, Filler K et al (2014) Telomere length: a review of methods for measurement. Nurs Res 63(4):289–299
114. Alter BP, Baerlocher GM, Savage SA, Chanock SJ, Weksler BB, Willner JP et al (2007) Very short telomere length by flow fluorescence in situ hybridization identifies patients with dyskeratosis congenita. Blood 110(5):1439–1447
115. Alter BP, Giri N, Savage SA, Rosenberg PS (2009) Cancer in dyskeratosis congenita. Blood 113(26):6549–6557
116. Calado RT, Regal JA, Kleiner DE, Schrump DS, Peterson NR, Pons V et al (2009) A spectrum of severe familial liver disorders associate with telomerase mutations. PLoS One 4(11):e7926
117. Ikeda H, Aida J, Hatamochi A, Hamasaki Y, Izumiyama-Shimomura N, Nakamura K et al (2014) Quantitative fluorescence in situ hybridization measurement of telomere length in skin with/without sun exposure or actinic keratosis. Hum Pathol 45(3):473–480
118. Kawano Y, Ishikawa N, Aida J, Sanada Y, Izumiyama-Shimomura N, Nakamura K et al (2014) Q-FISH measurement of hepatocyte telomere lengths in donor liver and graft after pediatric living-donor liver transplantation: donor age affects telomere length sustainability. PLoS One 9(4):e93749
119. Sanada Y, Aida J, Kawano Y, Nakamura K, Shimomura N, Ishikawa N et al (2012) Hepatocellular telomere length in biliary atresia measured by Q-FISH. World J Surg 36(4):908–916
120. Baerlocher GM, Lansdorp PM (2004) Telomere length measurements using fluorescence in situ hybridization and flow cytometry. Methods Cell Biol 75:719–750
121. Izumiyama-Shimomura N, Nakamura K, Aida J, Ishikawa N, Kuroiwa M, Hiraishi N et al (2014) Short telomeres and chromosome instability prior to histologic malignant progression and cytogenetic aneuploidy in papillary urothelial neoplasms. Urol Oncol 32(2):135–145
122. Poon SS, Lansdorp PM (2001) Quantitative fluorescence in situ hybridization (Q-FISH). Curr Protoc Cell Biol (Chapter 18:Unit 18 4). doi: 10.1002/0471143030.cb1804s12
123. Poon SS, Lansdorp PM (2001) Measurements of telomere length on individual chromosomes by image cytometry. Methods Cell Biol 64:69–96
124. Rufer N, Dragowska W, Thornbury G, Roosnek E, Lansdorp PM (1998) Telomere length dynamics in human lymphocyte subpopulations measured by flow cytometry. Nat Biotechnol 16(8):743–747
125. Baird DM, Rowson J, Wynford-Thomas D, Kipling D (2003) Extensive allelic variation and ultrashort telomeres in senescent human cells. Nat Genet 33(2):203–207
126. Cawthon RM (2009) Telomere length measurement by a novel monochrome multiplex quantitative PCR method. Nucleic Acids Res 37(3):e21
127. Bendix L, Horn PB, Jensen UB, Rubelj I, Kolvraa S (2010) The load of short telomeres, estimated by a new method, Universal STELA, correlates with number of senescent cells. Aging Cell 9(3):383–397
128. Hills M, Lucke K, Chavez EA, Eaves CJ, Lansdorp PM (2009) Probing the mitotic history and developmental stage of hematopoietic cells using single telomere length analysis (STELA). Blood 113(23):5765–5775

129. Lan Q, Cawthon R, Shen M, Weinstein SJ, Virtamo J, Lim U et al (2009) A prospective study of telomere length measured by monochrome multiplex quantitative PCR and risk of non-Hodgkin lymphoma. Clin Cancer Res 15(23):7429–7433
130. Shen M, Cawthon R, Rothman N, Weinstein SJ, Virtamo J, Hosgood HD 3rd et al (2011) A prospective study of telomere length measured by monochrome multiplex quantitative PCR and risk of lung cancer. Lung Cancer 73(2):133–137
131. de Lange T, Shiue L, Myers RM, Cox DR, Naylor SL, Killery AM et al (1990) Structure and variability of human chromosome ends. Mol Cell Biol 10(2):518–527
132. Counter CM, Hirte HW, Bacchetti S, Harley CB (1994) Telomerase activity in human ovarian carcinoma. Proc Natl Acad Sci U S A 91(8):2900–2904
133. Kim NW, Piatyszek MA, Prowse KR, Harley CB, West MD, Ho PL et al (1994) Specific association of human telomerase activity with immortal cells and cancer. Science 266(5193):2011–2015
134. Piatyszek MA, Kim NW, Weinrich SL, Hiyama K, Hiyama E, Wright WE et al (1995) Detection of telomerase activity in human cells and tumors by a telomeric repeat amplification protocol (TRAP). Methods Cell Sci 17:1–15
135. Bryan TM, Englezou A, Dalla-Pozza L, Dunham MA, Reddel RR (1997) Evidence for an alternative mechanism for maintaining telomere length in human tumors and tumor-derived cell lines. Nat Med 3(11):1271–1274
136. Gagos S, Chiourea M, Christodoulidou A, Apostolou E, Raftopoulou C, Deustch S et al (2008) Pericentromeric instability and spontaneous emergence of human neoacrocentric and minute chromosomes in the alternative pathway of telomere lengthening. Cancer Res 68(19):8146–8155
137. Garcia-Cao M, O'Sullivan R, Peters AH, Jenuwein T, Blasco MA (2004) Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases. Nat Genet 36(1):94–99
138. Cesare AJ, Reddel RR (2008) Telomere uncapping and alternative lengthening of telomeres. Mech Ageing Dev 129(1-2):99–108
139. Cesare AJ, Reddel RR (2010) Alternative lengthening of telomeres: models, mechanisms and implications. Nat Rev Genet 11(5):319–330
140. Henson JD, Neumann AA, Yeager TR, Reddel RR (2002) Alternative lengthening of telomeres in mammalian cells. Oncogene 21(4):598–610
141. Henson JD, Reddel RR (2010) Assaying and investigating alternative lengthening of telomeres activity in human cells and cancers. FEBS Lett 584(17):3800–3811
142. Muntoni A, Reddel RR (2005) The first molecular details of ALT in human tumor cells. Hum Mol Genet (14 Spec No. 2:R191-6)
143. Reddel RR (2003) Alternative lengthening of telomeres, telomerase, and cancer. Cancer Lett 194(2):155–162
144. Reddel RR, Bryan TM, Colgin LM, Perrem KT, Yeager TR (2001) Alternative lengthening of telomeres in human cells. Radiat Res 155(1 Pt 2):194–200
145. Sakellariou D, Chiourea M, Raftopoulou C, Gagos S (2013) Alternative lengthening of telomeres: recurrent cytogenetic aberrations and chromosome stability under extreme telomere dysfunction. Neoplasia 15(11):1301–1313
146. Londono-Vallejo JA, Der-Sarkissian H, Cazes L, Bacchetti S, Reddel RR (2004) Alternative lengthening of telomeres is characterized by high rates of telomeric exchange. Cancer Res 64(7):2324–2327
147. Carbone R, Pearson M, Minucci S, Pelicci PG (2002) PML NBs associate with the hMre11 complex and p53 at sites of irradiation induced DNA damage. Oncogene 21(11):1633–1640
148. Lamarche BJ, Orazio NI, Weitzman MD (2010) The MRN complex in double-strand break repair and telomere maintenance. FEBS Lett 584(17):3682–3695
149. Mirzoeva OK, Petrini JH (2001) DNA damage-dependent nuclear dynamics of the Mre11 complex. Mol Cell Biol 21(1):281–288

150. Munch S, Weidtkamp-Peters S, Klement K, Grigaravicius P, Monajembashi S, Salomoni P et al (2014) The tumor suppressor PML specifically accumulates at RPA/Rad51-containing DNA damage repair foci but is nonessential for DNA damage-induced fibroblast senescence. Mol Cell Biol 34(10):1733–1746
151. Shen TH, Lin HK, Scaglioni PP, Yung TM, Pandolfi PP (2006) The mechanisms of PML-nuclear body formation. Mol Cell 24(3):331–339
152. Xu ZX, Timanova-Atanasova A, Zhao RX, Chang KS (2003) PML colocalizes with and stabilizes the DNA damage response protein TopBP1. Mol Cell Biol 23(12):4247–4256
153. Grobelny JV, Godwin AK, Broccoli D (2000) ALT-associated PML bodies are present in viable cells and are enriched in cells in the G(2)/M phase of the cell cycle. J Cell Sci 113(Pt 24):4577–4585
154. Plantinga MJ, Pascarelli KM, Merkel AS, Lazar AJ, von Mehren M, Lev D et al (2013) Telomerase suppresses formation of ALT-associated single-stranded telomeric C-circles. Mol Cancer Res 11(6):557–567
155. Draskovic I, Arnoult N, Steiner V, Bacchetti S, Lomonte P, Londono-Vallejo A (2009) Probing PML body function in ALT cells reveals spatiotemporal requirements for telomere recombination. Proc Natl Acad Sci U S A 106(37):15726–15731
156. Dejardin J, Kingston RE (2009) Purification of proteins associated with specific genomic Loci. Cell 136(1):175–186
157. Liu L, Bailey SM, Okuka M, Munoz P, Li C, Zhou L et al (2007) Telomere lengthening early in development. Nat Cell Biol 9(12):1436–1441
158. Wang F, Yin Y, Ye X, Liu K, Zhu H, Wang L et al (2012) Molecular insights into the heterogeneity of telomere reprogramming in induced pluripotent stem cells. Cell Res 22(4):757–768
159. Zalzman M, Falco G, Sharova LV, Nishiyama A, Thomas M, Lee SL et al (2010) Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature 464(7290):858–863
160. Blasco MA, Lee HW, Rizen M, Hanahan D, DePinho R, Greider CW (1997) Mouse models for the study of telomerase. Ciba Found Symp 211:160–170, discussion 70-6
161. Blasco MA (2005) Mice with bad ends: mouse models for the study of telomeres and telomerase in cancer and aging. EMBO J 24(6):1095–1103
162. Rudolph KL, Chang S, Lee HW, Blasco M, Gottlieb GJ, Greider C et al (1999) Longevity, stress response, and cancer in aging telomerase-deficient mice. Cell 96(5):701–712
163. Artandi SE, Alson S, Tietze MK, Sharpless NE, Ye S, Greenberg RA et al (2002) Constitutive telomerase expression promotes mammary carcinomas in aging mice. Proc Natl Acad Sci U S A 99(12):8191–8196
164. Canela A, Martin-Caballero J, Flores JM, Blasco MA (2004) Constitutive expression of tert in thymocytes leads to increased incidence and dissemination of T-cell lymphoma in Lck-Tert mice. Mol Cell Biol 24(10):4275–4293
165. Gonzalez-Suarez E, Samper E, Ramirez A, Flores JM, Martin-Caballero J, Jorcano JL et al (2001) Increased epidermal tumors and increased skin wound healing in transgenic mice overexpressing the catalytic subunit of telomerase, mTERT, in basal keratinocytes. EMBO J 20(11):2619–2630
166. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G et al (2008) Lung cancer susceptibility locus at 5p15.33. Nat Genet 40(12):1404–1406
167. Rafnar T, Sulem P, Stacey SN, Geller F, Gudmundsson J, Sigurdsson A et al (2009) Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. Nat Genet 41(2):221–227
168. Tomas-Loba A, Flores I, Fernandez-Marcos PJ, Cayuela ML, Maraver A, Tejera A et al (2008) Telomerase reverse transcriptase delays aging in cancer-resistant mice. Cell 135(4):609–622
169. Jaskelioff M, Muller FL, Paik JH, Thomas E, Jiang S, Adams AC et al (2011) Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. Nature 469(7328):102–106

170. Bernardes de Jesus B, Vera E, Schneeberger K, Tejera AM, Ayuso E, Bosch F et al (2012) Telomerase gene therapy in adult and old mice delays aging and increases longevity without increasing cancer. EMBO Mol Med 4(8):691–704
171. de Lange T (2005) Shelterin: the protein complex that shapes and safeguards human telomeres. Genes Dev 19(18):2100–2110
172. Lipps HJ, Rhodes D (2009) G-quadruplex structures: in vivo evidence and function. Trends Cell Biol 19(8):414–422
173. Oganesian L, Bryan TM (2007) Physiological relevance of telomeric G-quadruplex formation: a potential drug target. Bioessays 29(2):155–165
174. Sen D, Gilbert W (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature 334(6180):364–366
175. Huppert JL (2008) Hunting G-quadruplexes. Biochimie 90(8):1140–1148
176. Leonetti C, Scarsella M, Riggio G, Rizzo A, Salvati E, D'Incalci M et al (2008) G-quadruplex ligand RHPS4 potentiates the antitumor activity of camptothecins in preclinical models of solid tumors. Clin Cancer Res 14(22):7284–7291
177. Salvati E, Leonetti C, Rizzo A, Scarsella M, Mottolese M, Galati R et al (2007) Telomere damage induced by the G-quadruplex ligand RHPS4 has an antitumor effect. J Clin Invest 117(11):3236–3247
178. Schoeftner S, Blasco MA (2009) A 'higher order' of telomere regulation: telomere heterochromatin and telomeric RNAs. EMBO J 28(16):2323–2336
179. Chandra A, Hughes TR, Nugent CI, Lundblad V (2001) Cdc13 both positively and negatively regulates telomere replication. Genes Dev 15(4):404–414
180. Meier B, Driller L, Jaklin S, Feldmann HM (2001) New function of CDC13 in positive telomere length regulation. Mol Cell Biol 21(13):4233–4245
181. Wright JH, Gottschling DE, Zakian VA (1992) Saccharomyces telomeres assume a non-nucleosomal chromatin structure. Genes Dev 6(2):197–210
182. Hardy CF, Sussel L, Shore D (1992) A RAP1-interacting protein involved in transcriptional silencing and telomere length regulation. Genes Dev 6(5):801–814
183. Tham WH, Zakian VA (2002) Transcriptional silencing at Saccharomyces telomeres: implications for other organisms. Oncogene 21(4):512–521
184. Krauskopf A, Blackburn EH (1998) Rap1 protein regulates telomere turnover in yeast. Proc Natl Acad Sci U S A 95(21):12486–12491
185. Kyrion G, Boakye KA, Lustig AJ (1992) C-terminal truncation of RAP1 results in the deregulation of telomere size, stability, and function in Saccharomyces cerevisiae. Mol Cell Biol 12(11):5159–5173
186. Levy DL, Blackburn EH (2004) Counting of Rif1p and Rif2p on Saccharomyces cerevisiae telomeres regulates telomere length. Mol Cell Biol 24(24):10857–10867
187. Marcand S, Brevet V, Gilson E (1999) Progressive cis-inhibition of telomerase upon telomere elongation. EMBO J 18(12):3509–3519
188. Kanoh J, Ishikawa F (2001) spRap1 and spRif1, recruited to telomeres by Taz1, are essential for telomere function in fission yeast. Curr Biol 11(20):1624–1630
189. Adams IR, McLaren A (2004) Identification and characterisation of mRif1: a mouse telomere-associated protein highly expressed in germ cells and embryo-derived pluripotent stem cells. Dev Dyn 229(4):733–744
190. Hu G, Kim J, Xu Q, Leng Y, Orkin SH, Elledge SJ (2009) A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Dev 23(7):837–848
191. Buonomo SB, Wu Y, Ferguson D, de Lange T (2009) Mammalian Rif1 contributes to replication stress survival and homology-directed repair. J Cell Biol 187(3):385–398
192. Silverman J, Takai H, Buonomo SB, Eisenhaber F, de Lange T (2004) Human Rif1, ortholog of a yeast telomeric protein, is regulated by ATM and 53BP1 and functions in the S-phase checkpoint. Genes Dev 18(17):2108–2119
193. Xu L, Blackburn EH (2004) Human Rif1 protein binds aberrant telomeres and aligns along anaphase midzone microtubules. J Cell Biol 167(5):819–830

194. Dan J, Liu Y, Liu N, Chiourea M, Okuka M, Wu T et al (2014) Rif1 maintains telomere length homeostasis of ESCs by mediating heterochromatin silencing. Dev Cell 29(1):7–19

195. Court R, Chapman L, Fairall L, Rhodes D (2005) How the human telomeric proteins TRF1 and TRF2 recognize telomeric DNA: a view from high-resolution crystal structures. EMBO Rep 6(1):39–45

196. Hwang H, Buncher N, Opresko PL, Myong S (2012) POT1–TPP1 regulates telomeric overhang structural dynamics. Structure 20(11):1872–1880

197. Griffith JD, Comeau L, Rosenfield S, Stansel RM, Bianchi A, Moss H et al (1999) Mammalian telomeres end in a large duplex loop. Cell 97(4):503–514

198. Gottschling DE, Zakian VA (1986) Telomere proteins: specific recognition and protection of the natural termini of Oxytricha macronuclear DNA. Cell 47(2):195–205

199. Greider CW (1999) Telomeres do D-loop-T-loop. Cell 97(4):419–422

200. de Lange T (2004) T-loops and the origin of telomeres. Nat Rev Mol Cell Biol 5(4): 323–329

201. Kuzminov A (2014) The precarious prokaryotic chromosome. J Bacteriol 196(10): 1793–1806

202. de Lange T, Petrini JH (2000) A new connection at human telomeres: association of the Mre11 complex with TRF2. Cold Spring Harb Symp Quant Biol 65:265–273

203. Zhu XD, Kuster B, Mann M, Petrini JH, de Lange T (2000) Cell-cycle-regulated association of RAD50/MRE11/NBS1 with TRF2 and human telomeres. Nat Genet 25(3):347–352

204. Brosh RM Jr (2013) DNA helicases involved in DNA repair and their roles in cancer. Nat Rev Cancer 13(8):542–558

205. Opresko PL, Otterlei M, Graakjaer J, Bruheim P, Dawut L, Kolvraa S et al (2004) The Werner syndrome helicase and exonuclease cooperate to resolve telomeric D loops in a manner regulated by TRF1 and TRF2. Mol Cell 14(6):763–774

206. Vannier JB, Pavicic-Kaltenbrunner V, Petalcorin MI, Ding H, Boulton SJ (2012) RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. Cell 149(4):795–806

207. Doksani Y, Wu JY, de Lange T, Zhuang X (2013) Super-resolution fluorescence imaging of telomeres reveals TRF2-dependent T-loop formation. Cell 155(2):345–356

208. Hockemeyer D, Daniels JP, Takai H, de Lange T (2006) Recent expansion of the telomeric complex in rodents: Two distinct POT1 proteins protect mouse telomeres. Cell 126(1):63–77

209. Martinez P, Blasco MA (2010) Role of shelterin in cancer and aging. Aging Cell 9(5):653–666

210. Wu L, Multani AS, He H, Cosme-Blanco W, Deng Y, Deng JM et al (2006) Pot1 deficiency initiates DNA damage checkpoint activation and aberrant homologous recombination at telomeres. Cell 126(1):49–62

211. Sfeir A, Kosiyatrakul ST, Hockemeyer D, MacRae SL, Karlseder J, Schildkraut CL et al (2009) Mammalian telomeres resemble fragile sites and require TRF1 for efficient replication. Cell 138(1):90–103

212. Sfeir A, de Lange T (2012) Removal of shelterin reveals the telomere end-protection problem. Science 336(6081):593–597

213. Kaul Z, Cesare AJ, Huschtscha LI, Neumann AA, Reddel RR (2012) Five dysfunctional telomeres predict onset of senescence in human cells. EMBO Rep 13(1):52–59

214. Takai H, Smogorzewska A, de Lange T (2003) DNA damage foci at dysfunctional telomeres. Curr Biol 13(17):1549–1556

215. Ballal RD, Saha T, Fan S, Haddad BR, Rosen EM (2009) BRCA1 localization to the telomere and its loss from the telomere in response to DNA damage. J Biol Chem 284(52):36083–36098

216. Carlos AR, Escandell JM, Kotsantis P, Suwaki N, Bouwman P, Badie S et al (2013) ARF triggers senescence in Brca2-deficient cells by altering the spectrum of p53 transcriptional targets. Nat Commun 4:2697

217. Dantzer F, Giraud-Panis MJ, Jaco I, Ame JC, Schultz I, Blasco M et al (2004) Functional interaction between poly(ADP-Ribose) polymerase 2 (PARP-2) and TRF2: PARP activity negatively regulates TRF2. Mol Cell Biol 24(4):1595–1607
218. Dynek JN, Smith S (2004) Resolution of sister telomere association is required for progression through mitosis. Science 304(5667):97–100
219. van Steensel B, de Lange T (1997) Control of telomere length by the human telomeric protein TRF1. Nature 385(6618):740–743
220. Loayza D, De Lange T (2003) POT1 as a terminal transducer of TRF1 telomere length control. Nature 423(6943):1013–1018
221. Paeschke K, Simonsson T, Postberg J, Rhodes D, Lipps HJ (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. Nat Struct Mol Biol 12(10):847–854
222. Torigoe H, Furukawa A (2007) Tetraplex structure of fission yeast telomeric DNA and unfolding of the tetraplex on the interaction with telomeric DNA binding protein Pot1. J Biochem 141(1):57–68
223. Zaug AJ, Podell ER, Cech TR (2005) Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension in vitro. Proc Natl Acad Sci U S A 102(31):10864–10869
224. Sfeir A, Kabir S, van Overbeek M, Celli GB, de Lange T (2010) Loss of Rap1 induces telomere recombination in the absence of NHEJ or a DNA damage signal. Science 327(5973):1657–1661
225. Karlseder J, Hoke K, Mirzoeva OK, Bakkenist C, Kastan MB, Petrini JH et al (2004) The telomeric protein TRF2 binds the ATM kinase and can inhibit the ATM-dependent DNA damage response. PLoS Biol 2(8):E240
226. Celli GB, Denchi EL, de Lange T (2006) Ku70 stimulates fusion of dysfunctional telomeres yet protects chromosome ends from homologous recombination. Nat Cell Biol 8(8):885–890
227. Blasco MA (2007) The epigenetic regulation of mammalian telomeres. Nat Rev Genet 8(4):299–309
228. Gottschling DE, Cech TR (1984) Chromatin structure of the molecular ends of Oxytricha macronuclear DNA: phased nucleosomes and a telomeric complex. Cell 38(2):501–510
229. Galati A, Magdinier F, Colasanti V, Bauwens S, Pinte S, Ricordy R et al (2012) TRF2 controls telomeric nucleosome organization in a cell cycle phase-dependent manner. PLoS One 7(4):e34386
230. Pisano S, Galati A, Cacchione S (2008) Telomeric nucleosomes: forgotten players at chromosome ends. Cell Mol Life Sci 65(22):3553–3563
231. Biessmann H, Mason JM (2003) Telomerase-independent mechanisms of telomere elongation. Cell Mol Life Sci 60(11):2325–2333
232. Roth CW, Kobeski F, Walter MF, Biessmann H (1997) Chromosome end elongation by recombination in the mosquito Anopheles gambiae. Mol Cell Biol 17(9):5176–5183
233. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM (1993) Transposons in place of telomeric repeats at a Drosophila telomere. Cell 75(6):1083–1093
234. Pardue ML, DeBaryshe PG (2003) Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet 37:485–511
235. Abad JP, Villasante A (1999) The 3′ non-coding region of the Drosophila melanogaster HeT-A telomeric retrotransposon contains sequences with propensity to form G-quadruplex DNA. FEBS Lett 453(1-2):59–62
236. Raffa GD, Ciapponi L, Cenci G, Gatti M (2011) Terminin: a protein complex that mediates epigenetic maintenance of Drosophila telomeres. Nucleus 2(5):383–391
237. Cenci G, Ciapponi L, Gatti M (2005) The mechanism of telomere protection: a comparison between Drosophila and humans. Chromosoma 114(3):135–145
238. Gao G, Walser JC, Beaucher ML, Morciano P, Wesolowska N, Chen J et al (2010) HipHop interacts with HOAP and HP1 to protect Drosophila telomeres in a sequence-independent manner. EMBO J 29(4):819–829

239. Raffa GD, Cenci G, Ciapponi L, Gatti M (2013) Organization and maintenance of Drosophila telomeres: the roles of terminin and non-terminin proteins. Tsitologiia 55(3):204–208
240. Raffa GD, Raimondo D, Sorino C, Cugusi S, Cenci G, Cacchione S et al (2010) Verrocchio, a Drosophila OB fold-containing protein, is a component of the terminin telomere-capping complex. Genes Dev 24(15):1596–1601
241. Arrighi FE, Hsu TC (1971) Localization of heterochromatin in human chromosomes. Cytogenetics 10(2):81–86
242. Chamla Y, Ruffie M (1976) Production of C and T bands in human mitotic chromosomes after heat treatment. Hum Genet 34(2):213–216
243. Dutrillaux B (1973) New system of chromosome banding: the T bands (author's transl). Chromosoma 41(4):395–402
244. Pardue ML, Gall JG (1970) Chromosomal localization of mouse satellite DNA. Science 168(3937):1356–1358
245. Scheres JM (1974) Production of C and T bands in human chromosomes after heat treatment at high pH and staining with “stains-all”. Humangenetik 23(4):311–314
246. Scheres JM (1976) CT banding of human chromosomes: description of the banding technique and some of its modifications. Hum Genet 31(3):293–307
247. Ferguson-Smith MA (2008) Cytogenetics and the evolution of medical genetics. Genet Med 10(8):553–559
248. Ferguson-Smith MA, Trifonov V (2007) Mammalian karyotype evolution. Nat Rev Genet 8(12):950–962
249. Mefford HC, Trask BJ (2002) The complex structure and dynamic evolution of human subtelomeres. Nat Rev Genet 3(2):91–102
250. Tommerup H, Dousmanis A, de Lange T (1994) Unusual chromatin in human telomeres. Mol Cell Biol 14(9):5777–5785
251. Gonzalo S, Jaco I, Fraga MF, Chen T, Li E, Esteller M et al (2006) DNA methyltransferases control telomere length and telomere recombination in mammalian cells. Nat Cell Biol 8(4):416–424
252. Bender CM, Pao MM, Jones PA (1998) Inhibition of DNA methylation by 5-aza-2′-deoxycytidine suppresses the growth of human tumor cell lines. Cancer Res 58(1):95–101
253. Dominguez-Bendala J, McWhir J (2004) Enhanced gene targeting frequency in ES cells with low genomic methylation levels. Transgenic Res 13(1):69–74
254. Shaffer LG, Lupski JR (2000) Molecular mechanisms for constitutional chromosomal rearrangements in humans. Annu Rev Genet 34:297–329
255. Ottaviani A, Gilson E, Magdinier F (2008) Telomeric position effect: from the yeast paradigm to human pathologies? Biochimie 90(1):93–107
256. Luke B, Lingner J (2009) TERRA: telomeric repeat-containing RNA. EMBO J 28(17): 2503–2510
257. Aparicio OM, Billington BL, Gottschling DE (1991) Modifiers of position effect are shared between telomeric and silent mating-type loci in S. cerevisiae. Cell 66(6):1279–1287
258. Stadler G, Rahimov F, King OD, Chen JC, Robin JD, Wagner KR et al (2013) Telomere position effect regulates DUX4 in human facioscapulohumeral muscular dystrophy. Nat Struct Mol Biol 20(6):671–678
259. Baur JA, Zou Y, Shay JW, Wright WE (2001) Telomere position effect in human cells. Science 292(5524):2075–2077
260. Renauld H, Aparicio OM, Zierath PD, Billington BL, Chhablani SK, Gottschling DE (1993) Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. Genes Dev 7(7A):1133–1145
261. Mondoux MA, Zakian VA (2007) Subtelomeric elements influence but do not determine silencing levels at Saccharomyces cerevisiae telomeres. Genetics 177(4):2541–2546
262. Horn D, Cross GA (1995) A developmentally regulated position effect at a telomeric locus in Trypanosoma brucei. Cell 83(4):555–561
263. Matzke MA, Moscone EA, Park YD, Papp I, Oberkofler H, Neuhuber F et al (1994) Inheritance and expression of a transgene insert in an aneuploid tobacco line. Mol Gen Genet 245(4):471–485

264. Nimmo ER, Cranston G, Allshire RC (1994) Telomere-associated chromosome breakage in fission yeast results in variegated expression of adjacent genes. EMBO J 13(16):3801–3811
265. Gehring WJ, Klemenz R, Weber U, Kloter U (1984) Functional analysis of the white gene of Drosophila by P-factor-mediated transformation. EMBO J 3(9):2077–2085
266. Hazelrigg T, Levis R, Rubin GM (1984) Transformation of white locus DNA in drosophila: dosage compensation, zeste interaction, and position effects. Cell 36(2):469–481
267. Levis R, Hazelrigg T, Rubin GM (1985) Effects of genomic position on the expression of transduced copies of the white gene of Drosophila. Science 229(4713):558–561
268. Gottschling DE, Aparicio OM, Billington BL, Zakian VA (1990) Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription. Cell 63(4):751–762
269. Boulton SJ, Jackson SP (1998) Components of the Ku-dependent non-homologous end-joining pathway are involved in telomeric length maintenance and telomeric silencing. EMBO J 17(6):1819–1828
270. Cooper JP, Nimmo ER, Allshire RC, Cech TR (1997) Regulation of telomere length and function by a Myb-domain protein in fission yeast. Nature 385(6618):744–747
271. Park MJ, Jang YK, Choi ES, Kim HS, Park SD (2002) Fission yeast Rap1 homolog is a telomere-specific silencing factor and interacts with Taz1p. Mol Cells 13(2):327–333
272. Tennen RI, Bua DJ, Wright WE, Chua KF (2011) SIRT6 is required for maintenance of telomere position effect in human cells. Nat Commun 2:433
273. Kaminker P, Plachot C, Kim SH, Chung P, Crippen D, Petersen OW et al (2005) Higher-order nuclear organization in growth arrest of human mammary epithelial cells: a novel role for telomere-associated protein TIN2. J Cell Sci 118(Pt 6):1321–1330
274. Netzer C, Rieger L, Brero A, Zhang CD, Hinzke M, Kohlhase J et al (2001) SALL1, the gene mutated in Townes-Brocks syndrome, encodes a transcriptional repressor which interacts with TRF1/PIN2 and localizes to pericentromeric heterochromatin. Hum Mol Genet 10(26):3017–3024
275. Koering CE, Pollice A, Zibella MP, Bauwens S, Puisieux A, Brunori M et al (2002) Human telomeric position effect is determined by chromosomal context and telomeric chromatin integrity. EMBO Rep 3(11):1055–1061
276. van der Maarel SM, Frants RR (2005) The D4Z4 repeat-mediated pathogenesis of facioscapulohumeral muscular dystrophy. Am J Hum Genet 76(3):375–386
277. Dixit M, Ansseau E, Tassin A, Winokur S, Shi R, Qian H et al (2007) DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. Proc Natl Acad Sci U S A 104(46):18157–18162
278. Gabellini D, Green MR, Tupler R (2004) When enough is enough: genetic diseases associated with transcriptional derepression. Curr Opin Genet Dev 14(3):301–307
279. Gabriels J, Beckers MC, Ding H, De Vriese A, Plaisance S, van der Maarel SM et al (1999) Nucleotide sequence of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. Gene 236(1):25–32
280. Lemmers RJ, van der Vliet PJ, Klooster R, Sacconi S, Camano P, Dauwerse JG et al (2010) A unifying genetic model for facioscapulohumeral muscular dystrophy. Science 329(5999):1650–1653
281. Lupski JR (2012) Digenic inheritance and Mendelian disease. Nat Genet 44(12):1291–1292
282. Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J (2007) Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. Science 318(5851):798–801
283. Azzalin CM, Lingner J (2015) Telomere functions grounding on TERRA firma. Trends Cell Biol 25(1):29–36
284. Luke B, Panza A, Redon S, Iglesias N, Li Z, Lingner J (2008) The Rat1p 5′ to 3′ exonuclease degrades telomeric repeat-containing RNA and promotes telomere elongation in Saccharomyces cerevisiae. Mol Cell 32(4):465–477
285. Schoeftner S, Blasco MA (2008) Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat Cell Biol 10(2):228–236
286. Martadinata H, Phan AT (2009) Structure of propeller-type parallel-stranded RNA G-quadruplexes, formed by human telomeric RNA sequences in K+ solution. J Am Chem Soc 131(7):2570–2578

287. Randall A, Griffith JD (2009) Structure of long telomeric RNA transcripts: the G-rich RNA forms a compact repeating structure containing G-quartets. J Biol Chem 284(21): 13980–13986
288. Xu Y, Kimura T, Komiyama M (2008) Human telomere RNA and DNA form an intermolecular G-quadruplex. Nucleic Acids Symp Ser (Oxf) (52):169–170
289. Bah A, Wischnewski H, Shchepachev V, Azzalin CM (2012) The telomeric transcriptome of Schizosaccharomyces pombe. Nucleic Acids Res 40(7):2995–3005
290. Greenwood J, Cooper JP (2012) Non-coding telomeric and subtelomeric transcripts are differentially regulated by telomeric and heterochromatin assembly factors in fission yeast. Nucleic Acids Res 40(7):2956–2963
291. Vrbsky J, Akimcheva S, Watson JM, Turner TL, Daxinger L, Vyskot B et al (2010) siRNA-mediated methylation of Arabidopsis telomeres. PLoS Genet 6(6):e1000986
292. Azzalin CM, Lingner J (2008) Telomeres: the silence is broken. Cell Cycle 7(9):1161–1165
293. Porro A, Feuerhahn S, Reichenbach P, Lingner J (2010) Molecular dissection of telomeric repeat-containing RNA biogenesis unveils the presence of distinct and multiple regulatory pathways. Mol Cell Biol 30(20):4808–4817
294. Nergadze SG, Farnung BO, Wischnewski H, Khoriauli L, Vitelli V, Chawla R et al (2009) CpG-island promoters drive transcription of human telomeres. RNA 15(12):2186–2194
295. Caslini C, Connelly JA, Serna A, Broccoli D, Hess JL (2009) MLL associates with telomeres and regulates telomeric repeat-containing RNA transcription. Mol Cell Biol 29(16): 4519–4526
296. Porro A, Feuerhahn S, Lingner J (2014) TERRA-reinforced association of LSD1 with MRE11 promotes processing of uncapped telomeres. Cell Rep 6(4):765–776
297. Deng Z, Wang Z, Stong N, Plasschaert R, Moczan A, Chen HS et al (2012) A role for CTCF and cohesin in subtelomere chromatin organization, TERRA transcription, and telomere end protection. EMBO J 31(21):4165–4178
298. Iglesias N, Redon S, Pfeiffer V, Dees M, Lingner J, Luke B (2011) Subtelomeric repetitive elements determine TERRA regulation by Rap1/Rif and Rap1/Sir complexes in yeast. EMBO Rep 12(6):587–593
299. Skourti-Stathaki K, Proudfoot NJ (2014) A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev 28(13):1384–1396
300. Arora R, Lee Y, Wischnewski H, Brun CM, Schwarz T, Azzalin CM (2014) RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells. Nat Commun 5:5220
301. Horard B, Gilson E (2008) Telomeric RNA enters the game. Nat Cell Biol 10(2):113–115
302. Lorenzi LE, Bah A, Wischnewski H, Shchepachev V, Soneson C, Santagostino M et al (2015) Fission yeast Cactin restricts telomere transcription and elongation by controlling Rap1 levels. EMBO J 34(1):115–129
303. Chawla R, Redon S, Raftopoulou C, Wischnewski H, Gagos S, Azzalin CM (2011) Human UPF1 interacts with TPP1 and telomerase and sustains telomere leading-strand replication. EMBO J 30(19):4047–4058
304. Zinsser F (1906) Atrophia cutis reticularis cum pigmentione, dystrophia unguium et leukokeratosis oris. Ikonogr Dermatol 5:219–223
305. Drachtman RA, Alter BP (1995) Dyskeratosis congenita. Dermatol Clin 13(1):33–39
306. de la Fuente J, Dokal I (2007) Dyskeratosis congenita: advances in the understanding of the telomerase defect and the role of stem cell transplantation. Pediatr Transplant 11(6): 584–594
307. Dokal I (2006) Dyskeratosis congenita: a cancer prone syndrome associated with telomerase deficiency. Hematology EHA 2:29–35
308. Savage SA, Alter BP (2009) Dyskeratosis congenita. Hematol Oncol Clin North Am 23(2):215–231
309. Knight S, Vulliamy T, Copplestone A, Gluckman E, Mason P, Dokal I (1998) Dyskeratosis Congenita (DC) Registry: identification of new features of DC. Br J Haematol 103(4):990–996

310. Vulliamy TJ, Marrone A, Knight SW, Walne A, Mason PJ, Dokal I (2006) Mutations in dyskeratosis congenita: their impact on telomere length and the diversity of clinical presentation. Blood 107(7):2680–2685
311. Armanios M, Chen JL, Chang YP, Brodsky RA, Hawkins A, Griffin CA et al (2005) Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. Proc Natl Acad Sci U S A 102(44):15960–15964
312. Kirwan M, Dokal I (2009) Dyskeratosis congenita, stem cells and telomeres. Biochim Biophys Acta 1792(4):371–379
313. Vulliamy T, Marrone A, Szydlo R, Walne A, Mason PJ, Dokal I (2004) Disease anticipation is associated with progressive telomere shortening in families with dyskeratosis congenita due to mutations in TERC. Nat Genet 36(5):447–449
314. Vulliamy TJ, Walne A, Baskaradas A, Mason PJ, Marrone A, Dokal I (2005) Mutations in the reverse transcriptase component of telomerase (TERT) in patients with bone marrow failure. Blood Cells Mol Dis 34(3):257–263
315. Aubert G, Lansdorp PM (2008) Telomeres and aging. Physiol Rev 88(2):557–579
316. Heiss NS, Knight SW, Vulliamy TJ, Klauck SM, Wiemann S, Mason PJ et al (1998) X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. Nat Genet 19(1):32–38
317. Marrone A, Walne A, Tamary H, Masunari Y, Kirwan M, Beswick R et al (2007) Telomerase reverse-transcriptase homozygous mutations in autosomal recessive dyskeratosis congenita and Hoyeraal-Hreidarsson syndrome. Blood 110(13):4198–4205
318. Vulliamy T, Marrone A, Goldman F, Dearlove A, Bessler M, Mason PJ et al (2001) The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. Nature 413(6854):432–435
319. Vulliamy TJ, Kirwan MJ, Beswick R, Hossain U, Baqai C, Ratcliffe A et al (2011) Differences in disease severity but similar telomere lengths in genetic subgroups of patients with telomerase and shelterin mutations. PLoS One 6(9):e24383
320. Yamaguchi H, Calado RT, Ly H, Kajigaya S, Baerlocher GM, Chanock SJ et al (2005) Mutations in TERT, the gene for telomerase reverse transcriptase, in aplastic anemia. N Engl J Med 352(14):1413–1424
321. Tsakiri KD, Cronkhite JT, Kuan PJ, Xing C, Raghu G, Weissler JC et al (2007) Adult-onset pulmonary fibrosis caused by mutations in telomerase. Proc Natl Acad Sci U S A 104(18):7552–7557
322. Vulliamy T, Beswick R, Kirwan M, Marrone A, Digweed M, Walne A et al (2008) Mutations in the telomerase component NHP2 cause the premature ageing syndrome dyskeratosis congenita. Proc Natl Acad Sci U S A 105(23):8073–8078
323. Walne AJ, Vulliamy T, Marrone A, Beswick R, Kirwan M, Masunari Y et al (2007) Genetic heterogeneity in autosomal recessive dyskeratosis congenita with one subtype due to mutations in the telomerase-associated protein NOP10. Hum Mol Genet 16(13):1619–1629
324. Walne AJ, Vulliamy T, Beswick R, Kirwan M, Dokal I (2008) TINF2 mutations result in very short telomeres: analysis of a large cohort of patients with dyskeratosis congenita and related bone marrow failure syndromes. Blood 112(9):3594–3600
325. Savage SA, Giri N, Baerlocher GM, Orr N, Lansdorp PM, Alter BP (2008) TINF2, a component of the shelterin telomere protection complex, is mutated in dyskeratosis congenita. Am J Hum Genet 82(2):501–509
326. Aalfs CM, van den Berg H, Barth PG, Hennekam RC (1995) The Hoyeraal-Hreidarsson syndrome: the fourth case of a separate entity with prenatal growth retardation, progressive pancytopenia and cerebellar hypoplasia. Eur J Pediatr 154(4):304–308
327. Kajtar P, Mehes K (1994) Bilateral coats retinopathy associated with aplastic anaemia and mild dyskeratotic signs. Am J Med Genet 49(4):374–377
328. Anderson BH, Kasher PR, Mayer J, Szynkiewicz M, Jenkinson EM, Bhaskar SS et al (2012) Mutations in CTC1, encoding conserved telomere maintenance component 1, cause Coats plus. Nat Genet 44(3):338–342

329. Polvi A, Linnankivi T, Kivela T, Herva R, Keating JP, Makitie O et al (2012) Mutations in CTC1, encoding the CTS telomere maintenance complex component 1, cause cerebroretinal microangiopathy with calcifications and cysts. Am J Hum Genet 90(3):540–549
330. Holohan B, Wright WE, Shay JW (2014) Cell biology of disease: telomeropathies: an emerging spectrum disorder. J Cell Biol 205(3):289–299
331. Conkright JJ, Na CL, Weaver TE (2002) Overexpression of surfactant protein-C mature peptide causes neonatal lethality in transgenic mice. Am J Respir Cell Mol Biol 26(1):85–90
332. Gross TJ, Hunninghake GW (2001) Idiopathic pulmonary fibrosis. N Engl J Med 345(7):517–525
333. Raghu G, Weycker D, Edelsberg J, Bradford WZ, Oster G (2006) Incidence and prevalence of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 174(7):810–816
334. Hodgson U, Laitinen T, Tukiainen P (2002) Nationwide prevalence of sporadic and familial idiopathic pulmonary fibrosis: evidence of founder effect among multiplex families in Finland. Thorax 57(4):338–342
335. Marshall RP, Puddicombe A, Cookson WO, Laurent GJ (2000) Adult familial cryptogenic fibrosing alveolitis in the United Kingdom. Thorax 55(2):143–146
336. Alder JK, Chen JJ, Lancaster L, Danoff S, Su SC, Cogan JD et al (2008) Short telomeres are a risk factor for idiopathic pulmonary fibrosis. Proc Natl Acad Sci U S A 105(35):13051–13056
337. Armanios MY, Chen JJ, Cogan JD, Alder JK, Ingersoll RG, Markin C et al (2007) Telomerase mutations in families with idiopathic pulmonary fibrosis. N Engl J Med 356(13):1317–1326
338. Cronkhite JT, Xing C, Raghu G, Chin KM, Torres F, Rosenblatt RL et al (2008) Telomere shortening in familial and sporadic pulmonary fibrosis. Am J Respir Crit Care Med 178(7):729–737
339. Lee HL, Ryu JH, Wittmer MH, Hartman TE, Lymp JF, Tazelaar HD et al (2005) Familial idiopathic pulmonary fibrosis: clinical features and outcome. Chest 127(6):2034–2041
340. Loyd JE (2003) Pulmonary fibrosis in families. Am J Respir Cell Mol Biol 29(3 Suppl):S47–S50
341. Parry EM, Alder JK, Qi X, Chen JJ, Armanios M (2011) Syndrome complex of bone marrow failure and pulmonary fibrosis predicts germline defects in telomerase. Blood 117(21):5607–5611
342. Steele MP, Speer MC, Loyd JE, Brown KK, Herron A, Slifer SH et al (2005) Clinical and pathologic features of familial interstitial pneumonia. Am J Respir Crit Care Med 172(9):1146–1152
343. Garcia CK, Wright WE, Shay JW (2007) Human diseases of telomerase dysfunction: insights into tissue aging. Nucleic Acids Res 35(22):7406–7416
344. Calado RT, Regal JA, Hills M, Yewdell WT, Dalmazzo LF, Zago MA et al (2009) Constitutional hypomorphic telomerase mutations in patients with acute myeloid leukemia. Proc Natl Acad Sci U S A 106(4):1187–1192
345. Kirwan M, Vulliamy T, Marrone A, Walne AJ, Beswick R, Hillmen P et al (2009) Defining the pathogenic role of telomerase mutations in myelodysplastic syndrome and acute myeloid leukemia. Hum Mutat 30(11):1567–1573
346. Marrone A, Stevens D, Vulliamy T, Dokal I, Mason PJ (2004) Heterozygous telomerase RNA mutations found in dyskeratosis congenita and aplastic anemia reduce telomerase activity via haploinsufficiency. Blood 104(13):3936–3942
347. Lyakhovich A, Ramirez MJ, Castellanos A, Castella M, Simons AM, Parvin JD et al (2011) Fanconi anemia protein FANCD2 inhibits TRF1 polyADP-ribosylation through tankyrase1-dependent manner. Genome Integr 2(1):4
348. Joksic I, Vujic D, Guc-Scekic M, Leskovac A, Petrovic S, Ojani M et al (2012) Dysfunctional telomeres in primary cells from Fanconi anemia FANCD2 patients. Genome Integr 3(1):6
349. Ghosh AK, Rossi ML, Singh DK, Dunn C, Ramamoorthy M, Croteau DL et al (2012) RECQL4, the protein mutated in Rothmund-Thomson syndrome, functions in telomere maintenance. J Biol Chem 287(1):196–209

350. Zhong ZH, Jiang WQ, Cesare AJ, Neumann AA, Wadhwa R, Reddel RR (2007) Disruption of telomere maintenance by depletion of the MRE11/RAD50/NBS1 complex in cells that use alternative lengthening of telomeres. J Biol Chem 282(40):29314–29322
351. Metcalfe JA, Parkhill J, Campbell L, Stacey M, Biggs P, Byrd PJ et al (1996) Accelerated telomere shortening in ataxia telangiectasia. Nat Genet 13(3):350–353
352. Feldser D, Strong MA, Greider CW (2006) Ataxia telangiectasia mutated (Atm) is not required for telomerase-mediated elongation of short telomeres. Proc Natl Acad Sci U S A 103(7):2249–2251
353. Hande MP, Balajee AS, Tchirkov A, Wynshaw-Boris A, Lansdorp PM (2001) Extra-chromosomal telomeric DNA in cells from Atm(-/-) mice and patients with ataxia-telangiectasia. Hum Mol Genet 10(5):519–528
354. Chu WK, Hickson ID (2009) RecQ helicases: multifunctional genome caretakers. Nat Rev Cancer 9(9):644–654
355. Payne M, Hickson ID (2009) Genomic instability and cancer: lessons from analysis of Bloom's syndrome. Biochem Soc Trans 37(Pt 3):553–559
356. Wyllie FS, Jones CJ, Skinner JW, Haughton MF, Wallis C, Wynford-Thomas D et al (2000) Telomerase prevents the accelerated cell ageing of Werner syndrome fibroblasts. Nat Genet 24(1):16–17
357. Opresko PL, Cheng WH, von Kobbe C, Harrigan JA, Bohr VA (2003) Werner syndrome and the function of the Werner protein; what they can teach us about the molecular aging process. Carcinogenesis 24(5):791–802
358. Crabbe L, Verdun RE, Haggblom CI, Karlseder J (2004) Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. Science 306(5703):1951–1953
359. Barefield C, Karlseder J (2012) The BLM helicase contributes to telomere maintenance through processing of late-replicating intermediate structures. Nucleic Acids Res 40(15):7358–7367
360. Ehrlich M, Buchanan KL, Tsien F, Jiang G, Sun B, Uicker W et al (2001) DNA methyltransferase 3B mutations linked to the ICF syndrome cause dysregulation of lymphogenesis genes. Hum Mol Genet 10(25):2917–2931
361. Hansen RS, Wijmenga C, Luo P, Stanek AM, Canfield TK, Weemaes CM et al (1999) The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. Proc Natl Acad Sci U S A 96(25):14412–14417
362. Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M et al (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature 402(6758):187–191
363. Yehezkel S, Segev Y, Viegas-Pequignot E, Skorecki K, Selig S (2008) Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. Hum Mol Genet 17(18):2776–2789
364. Yehezkel S, Shaked R, Sagie S, Berkovitz R, Shachar-Bener H, Segev Y et al (2013) Characterization and rescue of telomeric abnormalities in ICF syndrome type I fibroblasts. Front Oncol 3:35
365. Deng Z, Campbell AE, Lieberman PM (2010) TERRA, CpG methylation and telomere heterochromatin: lessons from ICF syndrome cells. Cell Cycle 9(1):69–74
366. Ji G, Ruan W, Liu K, Wang F, Sakellariou D, Chen J et al (2013) Telomere reprogramming and maintenance in porcine iPS cells. PLoS One 8(9):e74202
367. Bailey SM, Brenneman MA, Goodwin EH (2004) Frequent recombination in telomeric DNA may extend the proliferative life of telomerase-negative cells. Nucleic Acids Res 32(12):3743–3751
368. Morrish TA, Greider CW (2009) Short telomeres initiate telomere recombination in primary and tumor cells. PLoS Genet 5(1):e1000357

Chapter 11
Proximal Regulatory Elements with Emphasis on CpG Rich Regions

Pavlos Fanis

Introduction

All cells of a complex multicellular organism contain the same genome but carry out different biological processes such as proliferation, differentiation, cell survival and apoptosis. The execution of these diverse functions is characterized by differentially expression of the genes in the different cell types of the organism. The mammalian genome encodes 30,000–40,000 genes which are transcribed in the proper spatial and temporal patterns. The regulation of the genes can be controlled at many phases such as transcription initiation, elongation and termination. Moreover, the control of the transcription can be applied at other levels including RNA processing, export from the nucleus to the cytoplasm, mRNA translation and mRNA and protein degradation. To better understand the mechanisms that are responsible for the distinct gene expression patterns we need to have better knowledge of the transcriptional regulatory elements influencing the transcription. In the eukaryotic genome there are various types of transcriptional regulatory elements that are involved in the control of gene expression. Here, we discuss the general features of eukaryotic transcription focusing on the GC rich regions and their roles in health and disease. We give an outline of the eukaryotic transcription initiation process and then we focus at the structural, function and mechanistic function of GC rich regions in health and disease as well as methods currently used to identify transcription regulatory elements in the human genome.

P. Fanis, B.Sc., Ph.D. (✉)
Molecular Genetics Thalassaemia Department, The Cyprus Institute of Neurology
and Genetics, 6 International Airport Avenue, Ayios Dhometios, 2370 Nicosia, Cyprus
e-mail: pavlosf@cing.ac.cy

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_11

Eukaryotic Transcription

The initiation is the first step in transcription where the highest regulation is occurring. The process begins when the RNA polymerase (RNA pol) binds to template DNA strand and begins the production of the complementary RNA. In eukaryotes there are five RNA polymerases, RNA pol I that transcribes a set of genes that encode the ribosomal RNAs (rRNAs), RNA pol III that is responsible for the transcription of genes that encode the transfer RNA (tRNA), the 5S RNA and some small RNAs and RNA pol II (will be highlighted in this chapter) that transcribes the majority of genes that encode the mRNA which serve as template for the production of protein molecules [1–3]. Finally, in plants there are the RNA pol IV and the RNA pol V that synthesize small interfering RNA (siRNA) and RNAs involved in siRNA-directed heterochromatin formation, respectively [4, 5].

Protein-coding genes are controlled by *cis*-acting regulatory DNA elements and by *trans*-acting remodeler, mediator complexes and transcription factors that recognize the *cis*-acting regulatory DNA elements. The *cis*-acting regulatory elements can be divided into the distinct proximal regulatory elements (promoter) and distal regulatory elements (Fig. 11.1).

The basic transcription machinery is composed by the RNA pol II enzyme, the general transcription factors such as TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH and the mediator complex [6]. TFIID general transcription factor interact with the core promoter and participate in the recruitment of the pre-initiation complex (PIC) [7]. TFIID is a multi-subunit protein complex containing the TATA-box binding protein (TBP) and the TBP-associated factors (TAFIIs). After the TFIID binds to the promoter the general transcription factors, RNA pol II and the mediator complex are recruited in a highly ordered fashion to form the PIC. When formation of PIC at the promoter is complete, transcription proceeds through a series of steps before the fully establishment of the transcription elongation RNA pol II complex [8].

The formation of PIC in the core promoter of a gene it serves to accurately initiate transcription. Transcriptional activity can be further activated through changes in the activities of transcription factors (co-activators) [9]. Transcription factors are usually DNA-binding proteins that affect the transcription of a gene positively or negatively by binding to DNA regulatory elements or by interactions with other proteins [10]. The recognition sites of transcription factors are specific and usually located upstream of the core promoter. The precise sequence of a transcription factor binding site can be important for the binding strength of a transcription factor, thus can have an impact in the levels of gene expression [11]. Transcription factors often work in groups or complexes that allow various levels of transcriptional control [11] and are classified in different categories according to their DNA-binding domains such us basic leucine zipper (bZIP), basic helix-loop-helix (bHLH), zinc finger, homeodomain and nuclear hormone factors [12].

Transcription factors bind to DNA and mark a gene for activation or repression through interactions with co-activators and co-repressors. Co-activators and co-repressors bind to transcription factors and function by recruiting other proteins

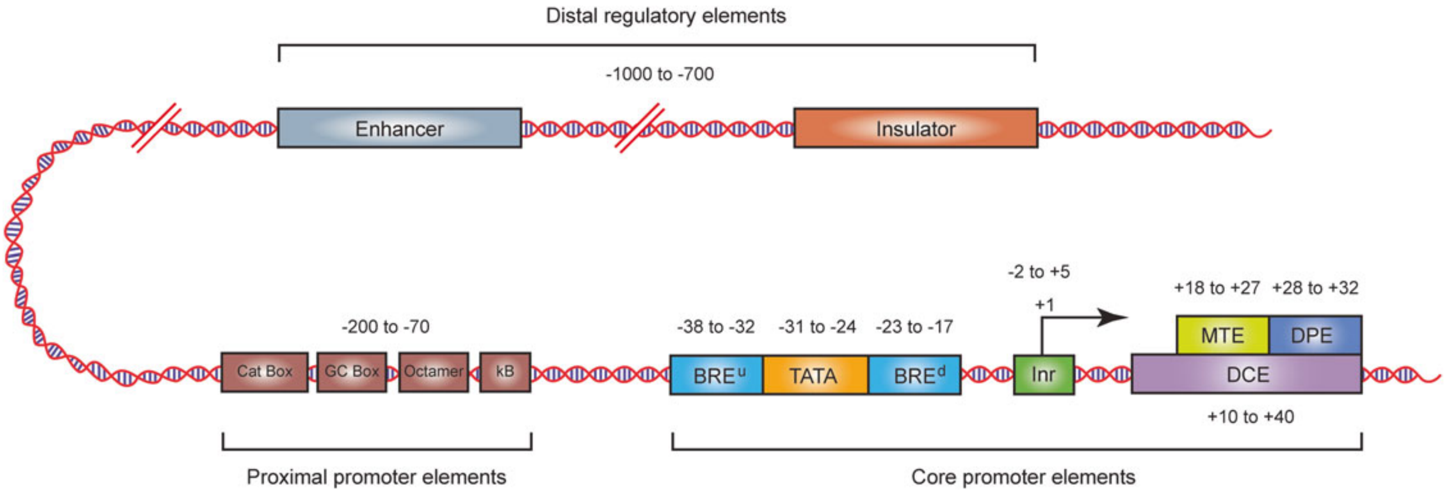


Fig. 11.1 Gene regulatory elements. Illustration of a typical eukaryotic gene regulatory region with clusters of core promoter elements, proximal promoter elements and distal regulatory elements. The promoter is composed of core promoter and proximal promoter elements. The core promoter elements may include the TATA box, the Initiator Element (Inr), the TFIIB-Recognition Element (BRE), the Downstream Promoter Element (DPE), the Motif Ten Element (MTE) and the Downstream Core Element (DCE). The Proximal promoter elements may include the CAAT box (Cat Box), the GC Box, the Octamer and the NF-kB (kB) elements. The locations of the elements are indicated relative to the transcription start site (+1)

with enzymatic activities that alter chromatin structure [13]. Chromatin compact state often makes gene promoters inaccessible to transcriptional machinery and/or transcription factors. Thus, co-activators including histone modification and ATP-dependent remodeling protein complexes alter the structure of chromatin by making it accessible to the binding of transcription factors [14]. The opposing effects on chromatin structure, by making it inaccessible to transcription factors, have the co-repressors.

Transcription in more than half of human genes is initiated from genomic regions with increase content of G and C nucleotides referred to as CpG islands.

GC Rich Regions

DNA methylation is carried out by the transfer of a methyl group at the 5 position of the pyrimidine ring of cytosine to form the methylcytosine and is observed in most of the organisms but the rate of methylation differs strongly, some species like yeast luck DNA methylation [15]. As the embryonic stem cell (ES) undergoes differentiation into different tissues, DNA methylation changes the expression of genes [16]. Methylation of DNA at the 5 position of cytosine has the specific effect of reducing gene expression. In somatic cells, DNA methylation usually occurs in a CpG dinucleotide context. CpG are DNA regions where a cytosine occurs next to a guanine. CpG means “Cytosine-phosphate-Guanine”. The CpG is used to distinguish the linear sequence from the G-C base pairing. Enzymes that add the methyl group to the cytosine are called DNA methyltransferases (DNMTs) [17].

The frequency of CpG dinucleotides in human genome is ~1 %, lower than would be expected due to random chance. It is proposed that the CpG deficiency is due to increase susceptibility of methylcytosines to deaminate to thymidine [18]. In mammals CpG islands are regions in the genome with a high content of CpG sites. CpG islands are typically 300–3000 bp in length and generally associated and found in or closed to ~40 % of promoter regions of mammalian genes [19, 20]. The CpG sites of the CpG island of promoters are unmethylated if the genes are expressed. Methylation of CpG sites in the promoter of a gene may inhibit gene expression [21]. The regulation of these promoters involves proteins which specifically bind at non-methylated CpGs and have an effect at the modification status of CpG island chromatin [22]. CpG islands typically exist at/or near transcription start site of genes, especially of housekeeping genes (genes that expressed in all cell types) and a significant fraction of the brain or the neutrally expressed genes [23, 24]. In humans, approximately 70 % of the gene promoters are associated with CpG islands making this the most common promoter type [25]. A large group of CpG islands are in distant regions from the transcription start site but show existence of promoter activity [20, 26].

Genes transcribed by RNA polymerase II can be divided in two different classes according to the CpG density across theirs 5′ ends. In the first class the CpG density is the same as the genome average (1 every 100 nucleotides). In this class there are genes that are expressed in a limited number of cell types. In the second class, the 5′ end of the genes is surrounded by CpG islands [27]. The association of CpG islands with the upstream regions of many genes can be used to predict promoters and/or genes in the mammalian genome [28]. As mentioned before, CpGs at the CpG islands of active genes remain not methylated. The features of non-methylation and the high density of G+C are accompanied by distinctive chromatin organization. The CpG islands show the properties of the active chromatin, such as hyperacetylation of histones H3 and H4 and nuclease enhance sensitivity at nucleosome-free regions [21, 29, 30].

CpG Islands and Methylation

CpG islands were first identified by digestion of mouse genomic DNA with a methyl-CpG specific restriction enzyme. A part of the DNA was consist of very highly fragmented DNA and was found to be composed of clusters of non-methylated CpG sites, the CpG islands [19, 45]. In addition, computational prediction and sequencing techniques identify approximately 27,000 CpG islands [43].

The characteristic clustering of CpG sites is because of their immunity against de novo methylation by DNA methyltransferases (DNMTs) during the earliest stages of mammalian development. A reason for this might be the binding of transcription factors that prohibit DNMT association at CpG island sequences [41].

The majority of CpG islands are hypomethylated but, as mentioned before, a small percentage is methylated during development. Some of these examples are shown to play a role in X-inactivation and genomic imprinting [46, 47]. Hypermethylation of CpG island promoters result in transcriptional repression. Promoters with relatively low CpG content were found to be more often hypermethylated [48]. Moreover, sites of CpG island methylation frequently found to genomic regions distal to promoters [49].

Transcription in de novo methylation CpG islands (in tumor cells or in cell lines) is strongly repressed. This event does not occur in physiological conditions in the organism except for the CpG islands of the imprinted genes [31, 32]. Repression of the transcription in the methylated DNA is mediated by proteins such as MeCP and MBD family proteins that bind specifically methylated CpGs and recruit histone deacetylases and transcriptional corepressors [33, 34]. DNA methylation in CpG-poor promoters correlates with their level of expression. Many examples suggest that although DNA methylation affects gene expression it is unlikely to play a general role as a transcriptional regulator. For example, demethylation of the promoter of the aminotransferase gene in rats does not lead to its activation in cells where it was previously methylated and inactive despite the presence of proteins that bind to the promoter [35]. DNA methylation is essential for proper mammalian development as shown by the embryonic lethality caused by disruption of the methyltransferase genes *Dnmt1*, *Dnmt3α* or *Dnmt3b* in mouse [36, 37].

CpG islands remain free of methylation in the heavily methylated genome. One possibility is the protection by the transcription factors that bind and provide less accessibility to DNA methyltransferases [41]. It is still unclear the existence of CpG islands because ~40 % of all human promoters are CpG-poor and can operate without CpG islands. An example is the erythroid specific α-and β-globin genes that are expressed to produce the haemoglobins. The α-globin gene is associated with a CpG island but the β-globin gene is not. The mouse α-and β-globin genes both are not associated with a CpG island [42]. This difference between human and mouse CpG islands can be found in approximately 20 % of human CpG island promoters compared to the mouse orthologues [23, 43] suggesting that some human CpG islands obtain a CpG island or some mouse promoters lost a CpG island during evolution [42]. CpG islands can be lost when they de novo methylated in the germ line and replaced through deamination by TpG instead of CpG but this cannot explained how they appear in the first place. One possibility is that CpG islands might have emerged as a genome footprint in the chromosomes by the replication initiation event [44]. Comparative foot-printing analyses across human and mouse CpG islands have revealed that the protein–DNA interaction pattern varies among the two species. Despite the high degree of conservation of their coding sequences this observation suggest that regulatory regions are more permissive to changes than coding regions [41].

Role of CpG Islands in Bidirectional Transcription and DNA Replivatio

An interesting feature of CpG island promoters is the elevated prevalence of bidirectional transcription [38]. The reason for this might be the high frequency of transcription factors bound to them and the presence of elements that are responsible for activating bidirectional transcription in vitro. This kind of organization allows the coordinate regulation of the two genes [39]. Some chromosomal replication origins have been mapped close to gene promoters in human cells. Moreover, it is also shown that transcription factors stimulate replication in many organisms. These evidences suggested that CpG islands might serve not only as promoters but also as replication origins because of the high density occupancy of transcription factors and the open chromatin organization. An example is the binding of the origin recognition complex (ORC) to the CpG islands [40].

Genomic Approaches for Identifying Regulatory Elements

There are many experimental and computational approaches for the identification or prediction of proximal regulatory elements in the genomes of eukaryotes.

Functional Assays

One of the most efficient ways for identifying and examining the regulatory activity of a DNA element is the use of a reporter gene assay. In such assay the DNA region of interest is cloned into a plasmid containing a reporter gene that can be easily measure [such as green fluorescent protein (GFP), luciferase, β-galactosidase, chloramphenicol acetyltransferase (CAT)]. The resulting construct is then introduced into cells or organism of interest and the reporter gene expression is measured. The nature of the plasmid construct depends on the regulatory elements to be identified. For example if the element to be tested is for promoter activity, is placed immediately upstream of the reporter gene. Once the element is identified further studies, such as deletions or creation of mutations, can be performed for more accurate characterization of the element. There are limitations of using such functional assays for identification of proximal regulatory elements. First, transfection assays mostly are performed in immortalized cell lines that are not representing the natural occurring environments. Secondly, an upstream regulatory element, in reality, might be used only in limited content such as specific tissue, developmental stage or specific environmental responses that differs from the cell culture that is selected for the assay. Transgenic assays by injecting the construct into embryos of animal models and follow the expression of the reporter gene through development can be done to overcome this limitation. Transgenic assays have also their limitation as they can reveal the specific expression pattern in the early developmental stages as they are sometimes instable because of embryonic cell multiplication [50].

Identification of Proximal Regulatory Elements on a Genome-Wide Scale

A technique for identification of transcription factor binding sites (TFBSs) on a genome-wide scale is the DNase I hypersensitive site (HS) mapping in which nucleosome-free regions are easily digested by the DNase I enzyme. These open chromatin regions are functionally related to transcriptional activity due to the binding of transcription factors [51, 52]. DNase I hypersensitive technique can be combined with high throughput sequencing or chip to provide a genome view of DNase I HS in specific cells at a specific developmental stage [53, 54].

Another powerful technique for determination of genomic sequences that are bound by a specific protein in vivo is the chromatin immunoprecipitation (ChIP). In this technique the protein(s) of interest are crosslinked temporarily with the associated chromatin in the living cells which then are lysed and the DNA–protein complexes are sheared by sonication at the desired size. The crosslinked DNA fragments that are associated with the protein(s) of interest are immunoprecipitated by a protein specific antibody and the associated DNA regions and their sequences can be determined by microarrays (ChIP-chip), by high throughput sequencing (ChIP-seq) or ChIP-exo, an extension of ChIP-seq to increase the resolution of TF bound sites [55]. Limitations of ChIP assays are that specific antibodies must be created for each DNA binding factor of interest. Moreover, ChIP assays cannot distinguish between different isoforms of a specific transcription factor.

Computational Approaches Identifying Proximal Regulatory Elements

Ninety eight percent of the human genome consist of non-coding DNA, thus is likely to contain regulatory regions [56]. Identifying a promoter of a specific gene can be difficult as core promoters are found far from the first exon because of the existence of the 5′ untranslated region (UTR) and/or introns. Furthermore, not all the promoters contain the core promoter elements thus the identification can be a challenge [57]. For prediction and identification of such regulatory regions, computational approaches are very helpful. Such approaches can look for common sequences to all known promoters and they search the genome to identify new regions with such sequences. These methods can be used alone or in combination with the existence of a CpG island or the presence of a possible first exon. Application of such approaches in genome-wide scale is limited because of lacking specificity and sensitivity. The reason of this is due to the fact that these computational programs rely on the amount and quality of the available data that they use for finding new regulatory regions/elements. Moreover, these approaches can identify proximal promoter elements that are already identified.

Identification of New Upstream Regulatory Elements

In TFBSs prediction programs a given sequence is scanned for known transcription factor sequence motifs that are experimentally identified. Examples of such databases are the TRANSFAC [58] and the JASPAR [59]. In these databases potential TFBSs are predicted, which in many cases the number of sites are large with many false positives. The reason for this is, in part, the quality of the data that are used to build the databases. Furthermore, these databases are not fully complete because not all of the DNA binding factors have been identified and some of the known DNA binding factors are not thoroughly defined. Another approach for identification of novel TFBSs is the examination for common sequence motifs in the upstream region of genes that are co-expressed. Known algorithms that use this approach are the MEME and AlignACE. Comparative genomic approaches can be used for prediction/identification of TFBSs. With these approaches TFBSs assumed to be conserved across evolution and DNA sequences from species separated by large evolutionary distances are compared. Sequences that are conserved are candidates to be functional TFBSs. Example of programs that perform such analyses are the PhastCons [60] and the Footprinter [61]. The limitation with these approaches is not all the TFBSs are common among species due to the fact that the same factor may bind to sequence variants of the TFBSs or a specific regulatory element is not conserved among species without affecting the expression of a gene. In addition, some regulatory elements important for human development and disease can be found only in humans. New analytical approaches in comparative genomics are required for detection of weak conserved TFBSs by increasing, for example, the number of species that the genome sequencing information is accessible.

References

1. Grummt I (1999) Regulation of mammalian ribosomal gene transcription by RNA polymerase I. Prog Nucleic Acid Res Mol Biol 62:109–154
2. Willis IM (1993) RNA polymerase III. Genes, factors and transcriptional specificity. Eur J Biochem 212:1–11
3. Sims RJ 3rd, Mandal SS, Reinberg D (2004) Recent highlights of RNA-polymerase-II-mediated transcription. Curr Opin Cell Biol 16:263–271
4. Herr AJ, Jensen MB, Dalmay T, Baulcombe DC (2005) RNA polymerase IV directs silencing of endogenous DNA. Science 308:118–120
5. Wierzbicki AT, Ream TS, Haag JR, Pikaard CS (2009) RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nat Genet 41:630–634
6. Thomas MC, Chiang CM (2006) The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol 41:105–178
7. He Y, Fang J, Taatjes DJ, Nogales E (2013) Structural visualization of key steps in human transcription initiation. Nature 495:481–486
8. Holstege FC, Fiedler U, Timmers HT (1997) Three transitions in the RNA polymerase II transcription complex during initiation. EMBO J 16:7468–7480
9. Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev 62:465–503
10. Karin M (1990) Too many transcription factors: positive and negative interactions. New Biol 2:126–131
11. Ptashne M, Gann A (1997) Transcriptional activation by recruitment. Nature 386:569–577
12. Pabo CO, Sauer RT (1992) Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem 61:1053–1095
13. Rosenfeld MG, Lunyak VV, Glass CK (2006) Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. Genes Dev 20:1405–1428
14. Spiegelman BM, Heinrich R (2004) Biological control through regulated transcriptional coactivators. Cell 119:157–167
15. Capuano F, Mulleder M, Kok R, Blom HJ, Ralser M (2014) Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. Anal Chem 86:3697–3702
16. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33 Suppl:245–254
17. Robertson KD, Uzvolgyi E, Liang G, Talmadge C, Sumegi J, Gonzales FA, Jones PA (1999) The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. Nucleic Acids Res 27:2291–2298
18. Chahwan R, Wontakal SN, Roa S (2010) Crosstalk between genetic and epigenetic information through cytosine deamination. Trends Genet 26:443–448

19. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40:91–99
20. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet 6:e1001134
21. Tazi J, Bird A (1990) Alternative chromatin structure at CpG islands. Cell 60:909–920
22. Blackledge NP, Klose R (2011) CpG island chromatin: a platform for gene regulation. Epigenetics 6:147–152
23. Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A 90:11995–11999
24. Gardiner-Garden M, Frommer M (1994) Transcripts and CpG islands associated with the pro-opiomelanocortin gene and other neurally expressed genes. J Mol Endocrinol 12:365–382
25. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103:1412–1417
26. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJ, Haussler D, Marra MA, Hirst M, Wang T, Costello JF (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466:253–257
27. Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209–213
28. Larsen F, Gundersen G, Lopez R, Prydz H (1992) CpG islands as gene markers in the human genome. Genomics 13:1095–1107
29. Antequera F, Macleod D, Bird AP (1989) Specific protection of methylated CpGs in mammalian nuclei. Cell 58:509–517
30. Gilbert SL, Sharp PA (1999) Promoter-specific hypoacetylation of X-inactivated genes. Proc Natl Acad Sci U S A 96:13825–13830
31. Antequera F, Boyes J, Bird A (1990) High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. Cell 62:503–514
32. Esteller M (2002) CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene 21:5427–5440
33. Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM, Eisenman RN, Bird A (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393:386–389
34. Wade PA (2001) Methyl CpG-binding proteins and transcriptional repression. Bioessays 23:1131–1137
35. Weih F, Nitsch D, Reik A, Schutz G, Becker PB (1991) Analysis of CpG methylation and genomic footprinting at the tyrosine aminotransferase gene: DNA methylation alone is not sufficient to prevent protein binding in vivo. EMBO J 10:2559–2567
36. Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69:915–926
37. Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99:247–257
38. Adachi N, Lieber MR (2002) Bidirectional gene organization: a common architectural feature of the human genome. Cell 109:807–809
39. Somma MP, Pisano C, Lavia P (1991) The housekeeping promoter from the mouse CpG island HTF9 contains multiple protein-binding elements that are functionally redundant. Nucleic Acids Res 19:2817–2824
40. Ladenburger EM, Keller C, Knippers R (2002) Identification of a binding region for human origin recognition complex proteins 1 and 2 that coincides with an origin of DNA replication. Mol Cell Biol 22:1036–1048
41. Cuadrado M, Sacristan M, Antequera F (2001) Species-specific organization of CpG island promoters at mammalian homologous genes. EMBO Rep 2:586–592

42. Antequera F (2003) Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60:1647–1658
43. Mouse Genome Sequencing, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
44. Antequera F, Bird A (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. Curr Biol 9:R661–R667
45. Cooper DN, Taggart MH, Bird AP (1983) Unmethylated domains in vertebrate DNA. Nucleic Acids Res 11:647–658
46. Reik W (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447:425–432
47. Edwards CA, Ferguson-Smith AC (2007) Mechanisms regulating imprinted genes in clusters. Curr Opin Cell Biol 19:281–289
48. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39:457–466
49. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 38:1378–1385
50. Jiang T, Xing B, Rao J (2008) Recent developments of biological reporter technology for detecting gene expression. Biotechnol Genet Eng Rev 25:41–75
51. Brenowitz M, Senear DF, Shea MA, Ackers GK (1986) Quantitative DNase footprint titration: a method for studying protein-DNA interactions. Methods Enzymol 130:132–181
52. Galas DJ, Schmitz A (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5:3157–3170
53. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16:123–131
54. Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, Prabhakar S (2013) Uniform, optimal signal processing of mapped deep-sequencing data. Nat Biotechnol 31:615–622
55. Collas P (2010) The current state of chromatin immunoprecipitation. Mol Biotechnol 45:87–100
56. Elgar G, Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet 24:344–352
57. Narlikar L, Ovcharenko I (2009) Identifying regulatory elements in eukaryotic genomes. Brief Funct Genomic Proteomic 8:215–230
58. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29:281–283

59. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32:D91–D94
60. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050
61. Blanchette M, Tompa M (2003) FootPrinter: a program designed for phylogenetic footprinting. Nucleic Acids Res 31:3840–3842

Chapter 12
Genomic Analysis Through High-Throughput Sequencing

Michalis Hadjithomas

Introduction

The advent of high-throughput sequencing in the early years of the millennium has disrupted the way research is being conducted in the biological sciences across the board, from human biology to microbial ecology. High-throughput sequencing was primarily developed to produce massive amounts of sequences in order to reconstruct the large genomes of higher-organisms, especially humans. However, it was quickly realized that massively parallel sequencing could be used to probe and answer questions that were previously outside the reach of scientific methods. This has led to the development of a multitude of high-throughput sequencing applications including exome sequencing (i.e. targeted re-sequencing of an organism's exons), long range chromatin interaction analysis, ribosome profiling, and many more, some of which we will discuss in this chapter.

The development of applications of high-throughput sequencing was paralleled by an explosion of technology development, which has caused the per-base cost of sequencing to drop dramatically, faster than was expected based on Moore's law, which predicts that the output of a technology doubles—and therefore its cost halves—every 2 years (Fig. 12.1) [1]. The natural consequence of this drop in costs is twofold; experiments can be done in a grander scale and smaller labs (thus more researchers) are able to apply this technology to their own research.

The purpose of this chapter is to give a high level overview of the most popular sequencing technologies in the market at the time of writing, and also introduce the reader to some of the major applications that can be used in the study of genomic

M. Hadjithomas, B.Sc., Ph.D. (✉)
DOE Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA 94598, USA

Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
e-mail: michalis.h@gmail.com

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution*, DOI 10.1007/978-1-4939-3070-8_12

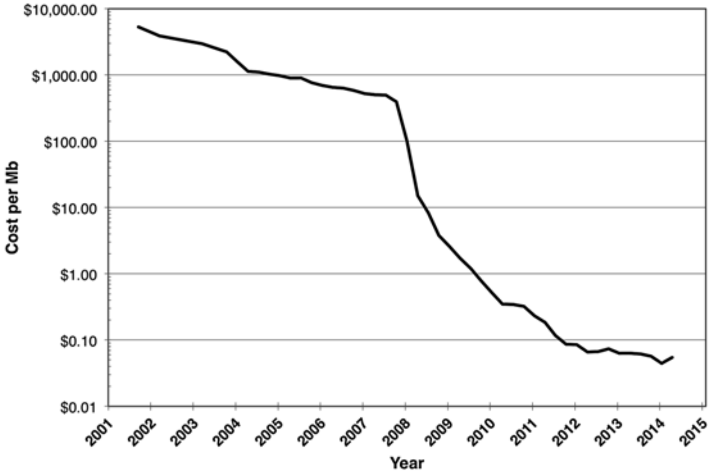


Fig. 12.1 The declining cost of DNA sequencing (data from NHGRI [1])

elements that play a role in human health. However, before starting the discussion, it is useful to define some of the terminologies that are often used when talking about high-throughput sequencing:

amplicon	An amplicon is the DNA fragment that is the product of PCR amplification of a target sequence.
contig	Most sequencing platforms produce short reads that are then assembled by computer algorithms to form longer sequences. A contig is a unit of the longest possible contiguous sequence that can be assembled by a specific algorithm using the available overlapping reads.
scaffold	A scaffold is a sequence that is composed of one or more contigs, which have been paired through the use of paired-end sequencing, or through the use of other experimental evidence, such as optical mapping.
read trimming	The computational operation to remove adapter sequences from reads after sequencing.
read	High-throughput sequencing is the result of the massive parallelization of individual sequencing reactions that are performed simultaneously. A "read" is the sequence output of each reaction.
polony	Polony derives from the term "polymerase colony" which describes a cluster of immobilized, clonally amplified DNA molecules.
paired-end tags	A paired-end tag (PET) is a fragment of DNA that results from the ligation of two sequences that may be on the opposite ends of a contiguous sequence or may be on different molecules but close in 3D space.
multiplexing	The combination of libraries coming from different samples in one sequencing reaction.
barcode	A short unique sequence of DNA added during library construction to allow for post-sequencing separation of reads coming from different sources/samples.

(continued)

(continued)

library	A collection of adapter ligated DNA fragments to be used as a template for sequencing.
coverage	Sometimes the term “coverage” is used inter-changeably with the term “depth of coverage”. However, when coverage is reported in percentages, it refers to the fraction of sequence recovered given a known length of input DNA. For example, if an amplicon is 1000 bases long, but only 900 bases are sequenced, the coverage is reported to be 90 %.
depth of coverage	The number of times that each nucleotide base is sequenced. The average depth of coverage can be calculated using Eq. (12.1) (“Overview and Costs in Applications” section below).

Technologies

There are three major players in the current market of high-throughput sequencers that are discussed in this section: Illumina/Solexa, Life Technologies’ Ion Torrent and Pacific Bioscience’s PacBio RSII system. Each technology has its strengths and weaknesses, which are summarized in Table 12.1. Naturally, the decision of which technology to acquire or use will depend on the intended application. A fourth emerging technology is not discussed here but is mentioned in the last section of the chapter, which discusses the future of this field.

A common characteristic of all the technologies discussed in this chapter is that they require a step of library construction before the sequencing run can be performed. The purpose of this step is to attach adapters to the ends of DNA molecules. For the Illumina and Ion Torrent platforms, these adapters are used for the clonal amplification of the library inserts in order to increase the detectable signal during sequencing. The sensitivity of the PacBio system is high enough that clonal amplification is not needed. Additionally, all these platforms need the adapters for sequencing initiation.

Illumina/Solexa

Illumina technology has the highest share of the high-throughput sequencing market as it is the oldest of the most popular technologies and also because its high sequencing output made it an attractive technology to acquire in the early days of high-throughput sequencing. Similarly to the other technology platforms discussed here, Illumina uses a Sequencing By Synthesis approach to elucidate the sequence of a DNA molecule.

The sequencing reactions in all Illumina platforms are performed on a flow cell, which may be divided from 1 to 8 lanes depending on the instrument being used. Attached to the flow cell are special oligonucleotides that are complimentary to the adapters used in the library preparation and are used to capture the single stranded DNA fragments. Using these complimentary adapter sequences, these fragments are clonally amplified clusters of DNA through bridge-PCR. During this process, the extended DNA forms a bridge with a near-by immobilized complimentary adapter,

Table 12.1 Summary of major sequencing technologies

Manufacturer	Version	Throughput	Read size	Strengths	Weaknesses
Illumina	MiSeq NextSeq 500 HiSeq	Up to 15 GB Up to 120 GB Up to 1 TB	Up to 300 bp Up to 150 bp Up to 250 bp	High throughput Established methodologies	Shorter read sizes
LifeTechnologies	PGM Proton	Up to 2 GB Up to 10 GB	Up to 400 bp Up to 200 bp	Less expensive technology to acquire Shortest run times	Errors associated with homopolymers
Pacific Biosciences	PacBio RS	Up to 1 GB	Up to 40 kbp reported	Long reads Direct detection of DNA modifications	Expensive Lower throughput

which now serves as the primer for the next round of synthesis. At the end of the clonal amplification reaction, one strand of DNA is removed from each template through the use of a restriction site on the flow cell bound adapter. The end result is a collection of single-stranded polonies that serve as the template for DNA polymerization.

Ion Torrent Semi-conductor Sequencing

The technology used by Life Technologies' IonTorrent platform is similar in many ways to the now defunct 454 pyrosequencing platform in that it uses emulsion-PCR (emPCR) for sample amplification and also in that only one type of nucleotide is provided during each flow cycle of the sequence run. IonTorrent semiconductor sequencing is unique in that it is the first mainstream platform that does not use light as the readout signal, but instead it detects the proton ions released by the addition of each nucleotide. The intensity of the signal is directly related to the number of bases incorporated at each cycle. A major advantage of this technique is that, since there is no need for image acquisition, sequencing run times can be as low as 2.3 h depending on the chip used. One drawback to this approach is that in cases where the same base is repeated multiple times (homopolymer) the signal output is high, which makes it challenging to accurately estimate the size of the repeat between homopolymers of similar size. For example, a homopolymer of nine cytosines may have a signal similar in intensity to that of a homopolymer of ten cytosines. This leads to a higher rate of false insertions or deletions in homopolymers.

There are two instruments offered by IonTorrent with different output capabilities, the Personal Genome Machine (PGM) and Proton. The PGM chip with the highest output (Ion318) produces up to 2 Gb of sequence per run, while the Proton can produce up to 10 Gb per run with the IonP1 chip. Life Technologies has announced recently that they will be releasing the Proton-II chip which will increase the output of the Proton to around 32 Gb which will enough to sequence a human genome to a depth of coverage of 10×, while maintaining the short sequencing run time. Additionally, Life Technologies is developing an alternative method to emPCR, which relies on isothermal template amplification [2]. The application of this method promises to greatly reduce template preparation time, without the need for a dedicated instrument, thus also reducing costs.

Pacific Biosciences SMRT

Pacific Biosciences (PacBio) with its single molecule real-time (SMRT) technology is the most recent player to emerge in the high-throughput sequencing field. The PacBio SMRT method relies on zero-mode waveguide (ZMW) wells, each containing a single polymerase enzyme, DNA template, sequencing primer and fluorescently labeled nucleotides. The fluorescent signal associated with each nucleotide incorporation to the growing DNA strand is recorded in real time [3]. This process takes about 2 days and it produces up to 5 Gb of sequence.

The PacBIO SMRT technology has three major strengths over the other sequencing methods discussed. First, it is the only technology, so far, that uses single molecule detection, which means it does not rely on PCR amplification that may introduce biases and errors. Secondly, it produces reads that are considerably longer that what Illumina and IonTorrent currently produce, with read lengths over 30,000 bases reported. These longer reads are extremely important when studying genomic regions with a high number of repetitive sequences. Additionally, longer reads allow for unambiguous mRNA isoform resolution since assembly of transcripts, and the artifacts associated with this process, is not necessary. Lastly, since DNA chain extension is observed in real time this method is able to capture polymerase kinetics. The rate with which polymerase incorporates nucleotides differs between modified and unmodified nucleotides. Consequently, the variation in the kinetics of nucleotide incorporation during DNA synthesis can be used to infer modifications in the DNA template and thus provide a direct way to studying base modifications [4].

Applications

Overview and Costs

In this section, major applications of high-throughput sequencing technologies will be discussed in some detail. However, this is by no means a complete collection of applications, since the number of applications is limited only by the number of methods that one can use to isolate DNA or RNA, and also because new applications will surely have emerged by the time this book is published. Software tools necessary for data analysis vary based on the application. Some commonly used and publicly available tools are summarized in Table 12.2, while other tools are specific to the sequencing technology and are provided by the manufacturer.

The general workflow of a high-throughput sequencing experiment is similar between most technologies in the market. The first step of the experiment, which in essence defines the application, is the isolation and selection of the nucleic acid material to be studied. These methodologies will be discussed briefly in each subsection. In the case of RNA based approaches, the RNA is reverse transcribed to DNA. Additionally, amplification of the isolated DNA is required for the Illumina and IonTorrent technologies. The resulting DNA molecules are then ligated to proprietary and technology specific adapters in the second step of library construction. The last step is the sequencing of the DNA libraries.

The cost of sequencing highly depends on different elements of the experimental design, in addition to the technology used. The first consideration is the depth of coverage of sequencing required, because this affects the amount of data (or output) that needs to be produced. The depth of coverage can be calculated using:

$$Depth\ of\ Coverage = \frac{O}{I} = \frac{R \times L}{I} \tag{12.1}$$

Table 12.2 Commonly used tools

	Name	Reference
Read Trimmers	Cutadapt	[42]
	PRINSEQ	[43]
	Skewer	[44]
	Trimmomatic	[45]
Genome assemblers	Celera	[46]
	Newbler	[47]
	SOAPdenovo2	[48]
	Velvet	[49]
	ALLPATHS	[50]
Transcriptomics	Tophat (Spliced alignment)	[51]
	Cufflinks (Assembler)	[52]
	PASA	[53]
	Rnnotator	[54]
	Trinity	[55]
Short read mappers	Bowtie	[56]
	BWA	[57]
	SOAP2	[58]

where O is the unassembled sequencing output, which can also be described as the number of reads (R) multiplied by the average length of each read (L), and I is either the estimated length of the input DNA (e.g. the size of a genome) or the sum of the length of the assembled contigs. Using this equation and having an estimate of the length of the DNA to be sequenced, a researcher may estimate the amount of sequencing needed to achieve the desired depth of coverage, which may be as high as the 1000× coverage required for confident identification of somatic mutations.

Often, the minimum output of a sequencing run may be orders of magnitude higher than the output needed for a sample, especially if the sample is comprised of only a limited number elements, as in the case of small RNAs. In these cases, the researcher may combine samples from multiple sources or experiments in one sequencing run by multiplexing. This is achieved through the addition of short unique DNA tags (barcodes) to each sample during library construction. The samples can then be separated computationally during post-sequencing analysis.

While multiple samples can be combined in one sequencing run, each sample will still need to go through the step of library construction. This raises the costs of sequencing experiments considerably since the economies of scale do not apply. Additionally, the DNA adapters and formulations necessary for library construction often are proprietary, which makes it difficult to find lower cost alternatives.

DNA Based Applications

Targeted Re-sequencing

Targeted re-sequencing is usually associated with sequencing the amplicons generated by the PCR amplification of exons. This can be performed on all predicted exons (i.e. the exome) or on a subset of these exons. This approach can easily be applied to the study of the non-genic genomic elements discussed in this book by simply designing custom PCR primers targeting areas of interest. Care needs to be taken when designing primers for targeted re-sequencing in order to avoid non-specific binding of oligo-nucleotides. In the case where a high number of loci are being targeted, potentially cross reacting primers maybe partitioned into separate primer mixes in order to be used in different PCR reactions.

An advantage of using this approach is that the length of the fragments to be sequenced are known in advance and therefore the step of fragment size selection can be avoided for established protocols. Additionally, knowing the input sequences simplifies downstream bioinformatic analyses.

Epigenomics

Epigenomics refers to the field that studies the global changes of gene expression that do not result from the change in DNA sequence, but from chemical modifications of either the DNA or the histones associated with DNA [5]. Epigenomic applications, therefore, depend on which of these two kinds of modifications one intends to study.

Chip-seq is used for the study of histone modifications. Chromatin, i.e. the complex of DNA and proteins, is crosslinked to form stable complexes, and then enzymatic treatment or physical shearing is used to break the chromatin into smaller pieces. DNA fragments bound to histones are selected using antibodies that target specific post-translational modifications. These fragments are sequenced and are then mapped to a reference genome to create a map of the particular modifications across the genome [6].

The methylation of DNA can be studied using a variety of approaches. The more straightforward approach is to use the Pacific Biosciences SMRT technology, which does not use PCR amplification and therefore preserves DNA modifications. The sequencing signal for this technology is the rate of base incorporation by polymerase during chain elongation. This rate depends on the chemical structure of the base being incorporated and therefore, also differs for chemically modified bases such as methylated adenines [7] and cytosines [8]. Additionally, this approach has the benefit that it does not require the existence of a reference sequence.

A more indirect way of studying DNA methylation of cytosines is using bisulfite sequencing. Bisulfite converts cytosine to uracil in treated DNA; however, methylated cytosines are protected from this conversion [9]. PCR amplification of bisulfite treated DNA will therefore result in the unmethylated cytosines to be converted to thymines while methylated cytosines will remain cytosines. The reads produced after a high-throughput sequencing run can then be mapped and compared to a reference genome to find which cytosines remained cytosines after bisulfite treatment, and were therefore methylated [10, 11].

Long-Distance Genomic Interactions

An innovative application of massively parallel high-throughput sequencing is the study of long distance chromatin interactions that occur between genomic elements. There are several variations of methods for studying the three-dimensional structure of chromatin in the nucleus, such as Circularized Chromosome Conformation Capture (4C) [12], Carbon-Copy Chromosome Conformation Capture (5C) [13], Hi-C [14], ChIA-PET [15] and others. All of these techniques are variations of the Chromosome Conformation Capture (3C) method in which chromatin is cross-linked, sheared (either enzymatically or physically) and then ligated. The result of this approach is that DNA fragments that are close in three-dimensional space but maybe far apart in sequence space or even on a different chromosome are linked together. Sequencing the ligated fragments, therefore, allows for studying the spatial organization of chromatin and how it differs between healthy and disease states, e.g. in healthy vs cancer cells [16].

Metagenomics

The term "metagenomics" describes the application of high-throughput sequencing in the study of microbial communities. It has traditionally been employed in the study of environmental samples, such as those gathered from soil or water, or samples collected from symbiotic environments, e.g. the root systems of plants. Metagenomics has more recently gained considerable traction in the study of human health. As a result of the colonization of intestinal and air-accessible surfaces on humans by bacteria and fungi, there are more non-human cells in the human body than human cells. It was therefore not a surprise that the composition of these microbial populations has a direct effect on human health. Some of the most impressive examples include the observation that the health of the gut microbiome is associated with diseases like obesity [17] and diabetes [18], or that the method of delivery at birth (viz. natural vs Caesarian) influences the health of the baby by changing the skin microbiome [19]. These studies, in addition to many others, employed extensive use of high-throughput sequencing to study the microbial communities inhabiting a human biome.

There are two common approaches to studying microbial communities using high-throughput sequencing. The first one targets the 16S ribosomal RNA genes (rRNA) of microbes using targeted PCR amplification. This approach gives an overview of the taxonomic composition of a microbiome. The main benefit of this approach is that it requires less sequencing throughput and the data analysis is relatively straight forward. However, preferential amplification of 16S rRNA genes may lead to over-representation of some bacterial groups [20]. Additionally, sequencing and assembly errors may lead to the formation of hard to identify chimeras [21]. The second approach, shotgun metagenomic sequencing, does not focus on specific genetic elements, but instead it is based on the genomic analysis of the total DNA extracted from microbial communities. Although the analysis of these data is more complex compared to 16S rRNA data, shotgun metagenomic sequencing is a much richer source of information. Besides revealing the composition of a microbiome, shotgun metagenomic sequencing can be used to study the functional properties of these organisms and to investigate the effect that the microbiome's secondary metabolism may have on the host [22]. Additionally, full reconstruction of microbial genomes is possible using these data [23]. Since shotgun metagenomic sequencing is unbiased, it can also be useful in the discovery of new bacterial groups and viruses [24].

RNA Based Applications

Small Non-coding RNA (miRNA) and Long Non-coding RNA (lncRNA) Sequencing

Non-coding RNAs play important roles in modulating the regulation of gene expression in all major cellular functions and, as such, they have been implicated in many diseases ranging from cancer to cardiac failure [25–27]. A major obstacle in studying ncRNAs with high-throughput sequencing methods stems from the difficulty in extraction and isolation of high quality samples at satisfactory yields, especially in the case of miRNAs (~22 bp in length). Up to 1 μg of miRNA can be isolated from cell lines and fresh tissue, but in the latter case the sample is heterogeneous [28]. Homogeneous samples from tissues can be retrieved through laser capture microdissection, which, however, drastically reduces the yield to less than 10 ng. Plasma and urine samples also have similarly low yields, which hinders accurate mRNA quantification and subsequent analyses. This problem is further exacerbated by the fact that miRNAs usually represent a small fraction of a cell's total RNA and the absence of a "hook" sequence that could be used for the targeted amplification or selection of these molecules.

Once an ncRNA sample has been attained, high-throughput sequencing can be used for the discovery of new ncRNA species [29] that could be used both to study their function but also as biomarkers for disease [30]. Additionally, using immuneprecipitation techniques (see below) the interactions of ncRNAs with other macromolecules can be further investigated.

RNA Interactions

Several methods exist to probe the interaction between RNA and proteins. Similarly to ChIP approaches, these methods involve the crosslinking of RNA to bound proteins, which are then selected using specific antibodies. Some of the most commonly used techniques are CLIP-seq [31], individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation (iCLIP) [32] and Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) [33]. These techniques enable the characterization and study of post-transcriptional control and the mechanisms through which mis-regulation may lead to disease [34]. For example, CLIP-Seq has been used to study the binding patterns of the exon junction complex, which plays a major role in the post-transcriptional fate of mRNA [35]. Another study determined the transcriptome-wide binding sites and preferences of RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) [33]. Understanding these processes will help elucidate the link between genetic variations and disease, especially in the cases of synonymous mutations or non-genic mutations that disrupt the mechanisms of post-transcriptional gene regulation.

Ribosome Profiling

The number of copies of an mRNA transcript present in a cell generally correlates with the level of its encoded protein being produced. However, this is not always the case as the translation of specific mRNAs may be attenuated. Ribosome profiling identifies which mRNAs are actively being translated, by using an approach similar that used for studying RNA–protein interactions [36]. Ribosomes and mRNA are crosslinked and then the overhanging strands of mRNA are digested by ribonucleases. The protected RNA is released, reverse transcribed to DNA that is then sequenced and mapped back to the original genome.

Besides providing a global view of all active protein expression in the cell, ribosome profiling also provides valuable information regarding translation elongation rates, and initiation and pause sites. Additionally, ribosome profiling allows for the discovery of elements that encode small proteins that may be missed by conventional methods. Lastly, the role of upstream open reading frames in the regulation of gene expression can be studied using this approach.

TAIL-seq

A very recently developed application for the study of gene expression regulation through the control of mRNA stability is TAIL-seq [37]. This method specifically targets the 3′ of mRNA molecules. Briefly, as in other RNA methods, total RNA is rRNA-depleted, tagged with a biotinylated adaptor at the 3′ end and then partially digested. The resulting fragment is pulled down using streptavidin, modified with a 5′ adapter, gel-purified, reverse transcribed and sequenced.

Although normally the length estimation of long homopolymers—in this case poly(A) and poly(T)—through sequencing is challenging, the authors of this study noticed that there was a correlation between the fluorescence signal intensity and the location of the end of a poly(T) homopolymer. Using this observation and statistical methods, they were able to accurately calculate the genome-wide poly(A) length distribution of mRNA tails in HeLA and NIH 3T3 in addition to observing pervasive uridylation and guanylation of poly(A) tails [37].

Future and Challenges

Sequencing technologies and protocols are constantly being improved as all of the companies try to remain competitive. New chemistries that extend read length and improve quality are being introduced regularly, whereas there is a push to simplify workflows as much as possible to make this technology accessible to any research or clinical lab, irrespective of its size. Illumina has recently announced that they will release model HiSeq X Ten in a bid to be the first technology to break the $1000 human genome barrier. The HiSeq X Ten system is in reality 10 HiSeq X instruments bundled together for a cost expected to exceed 10 million dollars. The major technological improvement in this new system is that the flow cells now come with ordered ‘nanowells’ where the template oligos are attached, as opposed to the random placement in the earlier systems. Additionally, Illumina is introducing a medium sized sequencer, NextSeq to occupy the space between MiSeq and HiSeq currently filled by Ion Torrent’s Proton sequencer.

Ion Torrent is also expected to release improved chemistries and protocols. The Ion Rapid Isothermal Amplification Chemistry is being developed to simplify and speed up library template preparation, which is now performed by the Ion Touch system. Additionally, Ion Torrent will release Ion Hi-Q™ Sequencing Chemistry, which is expected to reduce insertion and deletions errors by 90 %.

A lot of excitement has been created by the announcement of Oxford Nanopore Technologies’ intention to release two new sequencing systems, MinION and GridION, that use an entirely novel approach to sequencing. A biological nanopore is set in a membrane layer and allows electrical current to pass through. When an electrically charged particle, e.g. a nucleic acid polymer, passes through the pore it disrupts the electrical current. This disruption of current can be measured. Moreover, the intensity of signal change varies according to which base is passing through the nanopore, thus providing the means for DNA sequencing. The promise of this technology is the reduction both in sequencing costs but also the size of the sequencing instrument. For example, Oxford Nanopore’s MinION is a disposable device that is not much bigger than a pack of chewing gum and that connects directly to the USB port of a computer. The performance of MinION device sequencing is currently being evaluated through a community access program.

There has been considerable effort in recent years to develop the methodologies for applying genomic and transcriptomic techniques at the single cell level. The greatest challenge to this single cell sequencing is in amplifying the whole genome or transcriptome without introducing errors while avoiding contamination from other cells. Currently, the prevailing approaches to whole genome amplification are Multiple Displacement Amplification (MDA) [38] and Polymerase Chain Reaction (PCR). Both approaches have biases and may also lead to the introduction of artifacts; therefore, the decision of which method to use depends on the type of genetic elements and sequence variations to be studied [39]. Despite these hurdles, single cell sequencing will provide scientists with the capability to study the genomes and transcriptomes of healthy and diseased cells from the same individual over time, in addition to providing insights in the genomic heterogeneity of individuals. Lastly, a natural application of single cell genomics is in the field of pre-implantation genetic diagnosis [40] in addition to non-invasive pre-natal testing using isolated fetal cells from maternal blood [41].

With the continuing increase of sequencing throughput, the decrease in sequencing costs and the development of new technologies and methods there are two major hurdles for the adoption of high-throughput sequencing technologies by research labs. First is the availability of properly trained bioinformaticians who are able to analyze the large data sets produced by high-throughput sequencing. The second limitation is the cost associated with securely storing and managing these data, in addition to the cost of the high capacity computing capabilities needed to process and analyze the data. Despite these limitations, the benefits of acquiring high-throughput sequencing capabilities outweigh the associated effort and costs, as this opens new and unprecedented avenues towards understanding the role of non-genic elements in human health.

References

1. Wetterstand K, DNA Sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcosts/
2. Ma Z, Lee RW, Li B et al (2013) Isothermal amplification method for next-generation sequencing. Proc Natl Acad Sci 110:14320–14323. doi:10.1073/pnas.1311334110
3. Levene MJ, Korlach J, Turner SW et al (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299:682–686. doi:10.1126/science.1079700
4. Schadt EE, Banerjee O, Fang G et al (2013) Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res 23: 129–141. doi:10.1101/gr.136739.111
5. Callinan PA, Feinberg AP (2006) The emerging science of epigenomics. Hum Mol Genet 15:R95–R101. doi:10.1093/hmg/ddl095
6. Barski A, Cuddapah S, Cui K et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837. doi:10.1016/j.cell.2007.05.009
7. Fang G, Munera D, Friedman DI et al (2012) Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat Biotechnol 30:1232–1239. doi:10.1038/nbt.2432
8. Flusberg BA, Webster DR, Lee JH et al (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7:461–465. doi:10.1038/nmeth.1459
9. Frommer M, McDonald LE, Millar DS et al (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89:1827–1831
10. Colella S, Shen L, Baggerly KA et al (2003) Sensitive and quantitative universal Pyrosequencing methylation analysis of CpG sites. Biotechniques 35:146–150
11. Chatterjee A, Stockwell PA, Rodger EJ, Morison IM (2012) Comparison of alignment software for genome-wide bisulphite sequence data. Nucleic Acids Res 40(10):e79. doi:10.1093/nar/gks150
12. Zhao Z, Tavoosidana G, Sjölinder M et al (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 38:1341–1347. doi:10.1038/ng1891
13. Dostie J, Dekker J (2007) Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc 2:988–1002. doi:10.1038/nprot.2007.116
14. Belton J-M, McCord RP, Gibcus JH et al (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 58:268–276. doi:10.1016/j.ymeth.2012.05.001
15. Fullwood MJ, Liu MH, Pan YF et al (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462:58–64. doi:10.1038/nature08497
16. Fudenberg G, Getz G, Meyerson M, Mirny LA (2011) High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat Biotechnol 29:1109–1113. doi:10.1038/nbt.2049
17. Greenblum S, Turnbaugh PJ, Borenstein E (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A 109:594–599. doi:10.1073/pnas.1116053109

18. Qin J, Li Y, Cai Z et al (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490:55–60. doi:10.1038/nature11450
19. Dominguez-Bello MG, Costello EK, Contreras M et al (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci 107:11971–11975. doi:10.1073/pnas.1002601107
20. Mitra S, Förster-Fromme K, Damms-Machado A et al (2013) Analysis of the intestinal microbiota using SOLiD 16S rRNA gene sequencing and SOLiD shotgun sequencing. BMC Genomics 14:S16. doi:10.1186/1471-2164-14-S5-S16
21. Wylie KM, Truty RM, Sharpton TJ et al (2012) Novel bacterial taxa in the human microbiome. PLoS One 7:e35294. doi:10.1371/journal.pone.0035294
22. Donia MS, Cimermancic P, Schulze CJ et al (2014) A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158:1402–1414. doi:10.1016/j.cell.2014.08.032
23. Nielsen HB, Almeida M, Juncker AS et al (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822–828. doi:10.1038/nbt.2939
24. Yozwiak NL, Skewes-Cox P, Stenglein MD et al (2012) Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis 6:e1485. doi:10.1371/journal.pntd.0001485
25. Blume CJ, Hotz-Wagenblatt A, Hüllein J et al (2015) p53-dependent non-coding RNA networks in Chronic Lymphocytic Leukemia. Leukemia. doi:10.1038/leu.2015.119
26. Ardekani AM, Naeini MM (2010) The role of microRNAs in human diseases. Avicenna J Med Biotechnol 2:161–179
27. Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. Trends Cell Biol 21:354–361. doi:10.1016/j.tcb.2011.04.001
28. Pritchard CC, Cheng HH, Tewari M (2012) MicroRNA profiling: approaches and considerations. Nat Rev Genet 13:358–369. doi:10.1038/nrg3198
29. Williams Z, Ben-Dov IZ, Elias R et al (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. Proc Natl Acad Sci 110:4255–4260. doi:10.1073/pnas.1214046110
30. Jeffrey SS (2008) Cancer biomarker profiling with microRNAs. Nat Biotechnol 26:400–401. doi:10.1038/nbt0408-400
31. Murigneux V, Saulière J, Roest Crollius H, Le Hir H (2013) Transcriptome-wide identification of RNA binding sites by CLIP-seq. Methods 63:32–40. doi:10.1016/j.ymeth.2013.03.022
32. König J, Zarnack K, Rot G et al (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 17:909–915. doi:10.1038/nsmb.1838
33. Hafner M, Landthaler M, Burger L et al (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141:129–141. doi:10.1016/j.cell.2010.03.009
34. Clark PM, Loher P, Quann K et al (2014) Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. Sci Rep. doi:10.1038/srep05947
35. Saulière J, Murigneux V, Wang Z et al (2012) CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nat Struct Mol Biol 19:1124–1131. doi:10.1038/nsmb.2420
36. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223. doi:10.1126/science.1168978
37. Chang H, Lim J, Ha M, Kim VN (2014) TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. Mol Cell 53:1044–1052. doi:10.1016/j.molcel.2014.02.007
38. Spits C, Le Caignec C, De Rycke M et al (2006) Whole-genome multiple displacement amplification from single cells. Nat Protoc 1:1965–1970. doi:10.1038/nprot.2006.326
39. Macaulay IC, Voet T (2014) Single cell genomics: advances and future perspectives. PLoS Genet 10:e1004126. doi:10.1371/journal.pgen.1004126

40. Van der Aa N, Zamani Esteki M, Vermeesch JR, Voet T (2013) Preimplantation genetic diagnosis guided by single-cell genomics. Genome Med 5:71. doi:10.1186/gm475
41. Simpson JL (2013) Cell-free fetal DNA and maternal serum analytes for monitoring embryonic and fetal status. Fertil Steril 99:1124–1134. doi:10.1016/j.fertnstert.2013.02.012
42. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12. doi:http://dx.doi.org/10.14806/ej.17.1.200
43. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864. doi:10.1093/bioinformatics/btr026
44. Jiang H, Lei R, Ding S-W, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15:182. doi:10.1186/1471-2105-15-182
45. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl 30:2114–2120. doi:10.1093/bioinformatics/btu170
46. Miller JR, Delcher AL, Koren S et al (2008) Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24:2818–2824. doi:10.1093/bioinformatics/btn548
47. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327. doi:10.1016/j.ygeno.2010.03.001
48. Luo R, Liu B, Xie Y et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18. doi:10.1186/2047-217X-1-18
49. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. doi:10.1101/gr.074492.107
50. Maccallum I, Przybylski D, Gnerre S et al (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol 10:R103. doi:10.1186/gb-2009-10-10-r103
51. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. doi:10.1093/bioinformatics/btp120
52. Pollier J, Rombauts S, Goossens A (2013) Analysis of RNA-Seq data with TopHat and Cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures. Methods Mol Biol 1011:305–315. doi:10.1007/978-1-62703-414-2_24
53. Haas BJ, Salzberg SL, Zhu W et al (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9:R7. doi:10.1186/gb-2008-9-1-r7
54. Martin J, Bruno VM, Fang Z et al (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11:663. doi:10.1186/1471-2164-11-663
55. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. doi:10.1038/nprot.2013.084
56. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923
57. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324
58. Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinforma Oxf Engl 25:1966–1967. doi:10.1093/bioinformatics/btp336

Index

A

Adaptation, 9, 136, 208
Alternative lengthening of telomeres (ALTs), 254, 255, 257, 263
Aneuploidy, 145, 222, 232, 233, 235
Antagomirs, 19, 33, 35, 36

B

Bioinformatic, 7, 8, 19, 36, 163, 304
Biomarkers, 21, 27, 28, 30, 32, 33, 98, 100, 236, 306
Birth disorder, 222

C

Cancers, 11, 25–29, 32, 36, 48, 72, 84, 96–100, 115, 117, 130, 145–148, 165, 174, 175, 180, 181, 222, 227, 230, 232–236, 249, 252–255, 258, 262, 264, 266, 267, 305, 306
Centromere, 222–240, 254, 255, 260, 261, 265
Chromatin, 47, 48, 51, 54, 63, 64, 69, 84, 87–91, 94, 99, 101, 113–115, 162, 166–168, 172, 174, 223, 225, 227–229, 238, 258, 260–263, 288–291, 297, 304
Chromosome instability, 240
Chromosomes, 4, 54, 62–64, 67, 69, 85, 89, 90, 96, 97, 100, 101, 111, 113, 117, 131, 132, 136, 139, 140, 142, 157–160, 163, 164, 166, 168, 176, 179–181, 222–240, 247–251, 253–262, 264, 266, 267, 290, 305
Copy number variations (CNVs), 9, 11, 129–148, 160
CpG islands, 7, 72, 90, 114, 162, 167, 176, 235, 260, 262, 288–290, 292

D

Disease, 17, 21–25, 27–32, 34–36, 48, 85, 96, 98–100, 102, 113, 115, 116, 121, 129–131, 135, 136, 143–145, 147, 148, 161–163, 166, 167, 176, 179, 181, 226, 229, 232–236, 238, 261, 263, 264, 266, 267, 285, 293, 305, 306
DNA methylation, 48, 51, 53, 57, 61, 69, 71, 72, 90, 98, 113–115, 121, 235, 236, 260, 262, 288, 289, 304
DNA repair, 116, 147, 161, 167, 171, 173, 174, 204, 206, 207, 247, 249, 252, 257

E

Epigenetic, 28, 47–51, 53, 59, 63, 69, 72, 73, 89–91, 94, 99, 113, 142, 146, 160, 162, 222, 227, 230, 236, 238, 258, 260, 262
Epigenetic regulation, 90–91, 113, 116, 142
Epigenomics, 304
Evolution, 4–10, 59–65, 131–137, 143, 236–239

G

Gametogenesis, 48, 50, 52, 54, 58, 59, 62, 63, 66, 71–73, 171
Gene dosage, 9, 133, 141, 144
Gene expression, 1, 9, 11, 17, 32, 34, 48, 87–92, 97, 101, 112, 119, 121, 135, 137, 141, 142, 159, 161–163, 166, 167, 171, 174, 234, 285, 286, 288, 289, 291, 304, 306, 307

K. Felekkis, K. Voskarides (eds.), *Genomic Elements in Health, Disease and Evolution: Junk DNA*, DOI 10.1007/978-1-4939-3070-8

Gene expression regulation, 1, 9, 12, 88, 101, 307
Genome, 1, 2, 6, 7, 9–11, 19, 26, 35, 47, 48, 50, 54, 64, 69, 71, 73, 83, 84, 86–88, 91, 97, 100, 101, 109–113, 115, 117, 119, 121, 129–132, 135–138, 140–143, 146, 156–161, 163–170, 176, 180, 181, 198, 201, 206, 207, 209, 210, 225, 229, 234, 235, 237, 239, 240, 247–250, 261, 263, 267, 285, 288, 289, 291–293, 301, 303, 304, 307, 308
Genome-wide, 50, 84, 86–88, 91, 97, 100, 121, 129, 140, 141, 143, 182, 235, 267, 291, 292, 307
Germ cells, 48–55, 57–59, 61, 67–72, 115, 130, 166, 171–173, 175, 257
Group II introns, 195, 196, 198, 199, 202, 205–207, 209

H
Heterochromatic, 258–261
Heterochromatin, 47, 49, 63, 71, 90, 98, 113, 114, 160, 222, 228, 233, 235, 236, 254, 256, 260–262, 286
Heterochromatization, 113
High-throughput sequencing, 22, 297, 298
Human disease, 48, 71–72, 102, 116, 117, 148, 173–174, 229, 233, 250, 261
Human physiology and disease pathogenesis, 109–122

I
Immune-precipitation, 306
Immuno-precipitation, 22, 55, 262, 291, 306
Intellectual disability, 143, 148, 161
Interactome, (AU: Found only in the reference)
Inverted duplications, 6, 7, 136, 137

K
Kinetochore, 222, 223, 226–229, 231, 234, 238

L
Linkage, 23, 159, 176, 179, 182
Long non-coding RNA, 83, 84, 141

M
Maturases, 210
Medical genetics, 175, 176
Megasatellites, 156, 157, 159, 160, 166
Metagenomics, 305
MicroRNAs, 3, 22, 26, 31, 60, 84, 115, 136, 137, 141, 306
Microsatellites, 157, 158, 162–165, 168, 169, 174, 175, 179
Minisatellites, 156, 157, 159, 164–166, 168, 170, 175, 176, 180, 181
miRNA genes, 2–8, 10, 11, 22–24, 26, 28, 142
miRNA targets, 2, 9, 17–20, 24–26
miRSNPs, 18, 24–26
Mobile genetic elements, 198, 202, 210

N
Natural selection, 10, 11, 166, 167
Next-generation sequencing, 21
Non-pathogenic, 142

O
Oncogenes, 28, 142, 146, 232
Oncogenesis, 97

P
Paternity testing, 180
p-element induced wimpy testis proteins (PIWI), 48, 51–61, 63, 68–73, 115
piRNAs, 48, 49, 51–57, 61–73, 115
Promoters, 2, 86–88, 93, 94, 99, 112, 114, 115, 141, 161–163, 167, 208, 235, 262, 287–290, 292
Proximal regulatory elements, 286, 290, 291

R
Repeat expansion disorders, 165, 171, 172, 174, 175
Repeat instability, 161, 171–173
Repetitive DNA, 157, 159, 222
Retroelements, 109, 110, 116, 117, 170, 195, 210
Retrohoming, 202, 204–208
Reverse transcriptase (RTs), 47, 110, 140, 195–198, 204, 205, 210, 248–251, 253, 264–266
Ribozymes, 196, 197, 199, 201, 202, 207, 208, 248
RNA, 1–4, 17, 20–22, 31, 36, 47, 48, 50, 51, 53–57, 59, 62–67, 69, 70, 73, 83–87, 89, 90, 92–97, 99, 100, 102, 109, 110, 112–116, 118, 121, 141, 159, 162, 163, 169, 170, 172, 177, 178, 195, 196, 198, 200–202, 204, 206–209, 235, 248–252, 259, 262, 263, 285, 286, 288, 302, 305–307

RNA polymerases, 2, 83, 87, 110, 112, 119, 162, 208, 251, 262, 286, 288
RNA-seq, 84–87, 97, 200, 262

S
Satellite DNA, 156, 157, 159, 170, 224–226, 231, 236–239
Segmental duplications (SD), 6, 7, 131, 136, 137, 144, 260
Self-catalytic RNA
Shelterin, 257, 258, 264, 265
Splicing, 3, 31, 83, 86, 88, 89, 94–97, 112, 160, 164, 195–204, 206, 208–210, 235, 252
Syndrome, 23, 24, 98, 101, 130, 135, 142–144, 162, 175, 176, 179, 181, 230, 233, 235, 236, 250, 257, 258, 264–267

T
Tandem repeats (TRs), 156–170
Targetrons, 208, 209
Telomerases, 251–252, 254, 265
Telomere position effects (TPE), 260, 261
Telomeres, 157, 166, 247–263, 265, 266
Telomeric repeat-containing RNA (TERRA), 98, 99, 262, 263, 267
Telomeropathy, 263–267
Transcription, 2–4, 9, 17, 19–21, 28, 29, 33, 34, 48–50, 57–59, 61–64, 69, 71, 73, 84, 86–98, 101, 110, 112–115, 119–122, 132, 160–164, 167, 169, 170, 172, 198, 200, 204, 206, 229, 236, 252, 254, 260, 262, 263, 285–292, 306, 308
Transcriptomes, 5, 86, 101, 262, 267, 303, 306, 308
Transposable elements, 7, 47, 73, 85, 86, 112, 114, 116, 137, 169, 210, 225, 226, 235
Transposons, 7, 47, 53, 54, 57, 59, 62, 63, 65–67, 69, 73, 86, 109, 111, 116–118, 156, 169
Transposon silencing, 57, 59

U
Unveiling transposable elements function, 109–122

V
Variable number tandem repeats (VNTRs), 109, 110, 114, 156, 159, 165, 170, 180, 181