Chapter 5
The Mouse Genome

5.1 Introduction

In Chap. 4 we explained how mouse geneticists were able to develop high-density and high-resolution genetic maps of the mouse genome by taking advantage of the unequaled strategies and tools they had at their disposition: i.e., inter-subspecific crosses, recombinant inbred strains, radiation hybrids and a wealth of polymorphic molecular markers of all kinds. We also explained how the same geneticists could develop physical maps by anchoring virtual (i.e., in silico) DNA fragments cloned into BACs, YACs or cosmids onto the molecular markers previously ordered along each chromosome. It is clear that, while building these maps and associated libraries of cloned DNAs, geneticists were in fact gathering the essential ingredients for undertaking the logical next step: the sequencing of the whole mouse genome.

The decision to undertake such an ambitious (and, at the time, expensive) project was made at the turn of the millennium and was strongly influenced by the decision to sequence the human genome, made a few years earlier (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). A first draft of the mouse genome sequence was released in 2002, only a few months after the release of the first draft sequences of the human genome (Mural et al. 2002; Mouse Genome Sequencing Consortium, Waterston et al. 2002) and 2 years before the publication of the rat sequence (Gibbs et al. 2004).

The completion of these projects, as we will see in this chapter and the following chapters, had an enormous impact in many areas of genetics and biology. Making these genome sequences available to the community provided a wealth of information about genome structure and evolution through the identification of similarities and differences across species. As Robert Waterston and his colleagues wrote in the conclusion of their publication: "*The mouse provides a unique lens through which we can view ourselves* […]. *With the availability of* [its genome] *sequence, it* […] *provides a model and informs the study of our genome as well*" (Mouse Genome Sequencing Consortium, Waterston et al. 2002).

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_5

Nowadays, geneticists have direct and free access to a variety of high-quality genomic sequences through the Internet, and most of them would probably find it difficult to work without having these tools at hand.

5.2 The Sequence of the Mouse Genome

The availability of the mouse genome sequence represented such an important piece of information for the development of the genetics of this species that it would certainly have become available sooner or later, for example, as a consequence of the continuous addition of the ever-increasing number of sequence fragments released by independent laboratories. However, such a disorganized approach would have inevitably resulted in delay, in a sequence with plenty of gaps and redundancies, and finally in a higher cost. Retrospectively, the decision to give support and priority to the complete and systematic sequencing of the mouse genome, and to make it a concerted project completed by a team of specialists, should be considered very wise. This decision was also very altruistic because the laboratories that did not have easy access to sequencing facilities, for whatever reasons, can now benefit from this resource, entirely free of charge, for designing their experiments. Further evidence of this achievement is provided by the enormous and ever-increasing number of scientific papers that have been published since the release of the initial draft sequences of the mouse genome and make direct reference to it. This trend will certainly grow in the years to come with the progress made in sequencing technologies and the associated dramatic reduction in cost.

The sequence of the rat genome has also turned out to be a valuable piece of information for geneticists, because it has allowed three-way comparisons with the other two species (human and mouse). These comparisons have provided details about how evolution proceeds over a relatively short timescale. As mentioned in Chap. 1, the human and rodent lineages split around 75–80 Myr ago, while the mouse and rat lineages split around 12–14 Myr ago.

5.2.1 The Mouse Genome is Enormous in Size, and its Structure is Complex

Measurements of the intensity of the brilliant purple color performed on mouse cell nuclei (early spermatids, for example), after a Feulgen reaction, indicated that the DNA content of the mouse haploid genome corresponds to approximately 3×10^{-12} g (= 3 pg), which translates into a molecular weight of ~1.8×10^{12} daltons (Da). Since the average molecular weight of a double-stranded DNA base-pair (bp) is ~600 Da, this means that one expects to find ~3×10^9 bp or 3.0

Giga-base-pairs (Gb) of DNA in a mouse haploid genome (Silver 1995). This is ~650 times more than in the genome of the bacterium *Escherichia coli* K-12, which comprises 4,639,221 bp. To translate this into more palpable terms, we computed that, if the haploid mouse DNA sequence was printed as a single line using the 11-point `Courier` font, all in uppercase, to symbolize the four bases (`A`, `T`, `G`, `C`), the length of this line would be roughly equal the distance from London to New York City (5,600 km or 3,480 miles). To express this still differently, the printed transcription of the message in 12-point Times font would represent around 3,500 books with a size similar to the one you have in your hands. However, although obviously enormous, this sequence can be stored on the hard disk of a personal computer (Silver 1995). Finally, mouse nuclear DNA has an A + G/C + T ratio of 49.99/50.01 (~1), as in human.

Aside from its large size, the mouse genome is also heterogeneous. Biophysicists who studied the thermodynamics of nucleic acid denaturation/renaturation had already recognized this peculiarity, over 40 years ago, by measuring the $C_0t_{1/2}$ value, a parameter reflecting the structural heterogeneity of a DNA sample that is based on the speed of reconstitution of double-stranded DNA (dsDNA) from previously denatured single-stranded DNA (ssDNA). The same biophysicists also demonstrated that some fractions of the mouse genomic DNA renatured much faster than others as a consequence of a high proportion of repeated sequences.

Another interesting comparison is between the physical size of the mouse and pufferfish (*Takifugu rubripes*) genomes, leading to the observation that the genome of the fish is about nine times smaller (0.35 pg of DNA or 340 Mb) than that of the mouse. Considering that all vertebrates presumably have a similar number of protein-coding genes (between 20,000 and 30,000, as we will discuss further), it has been suggested that the difference in size between the two genomes is probably due to the presence, in the mouse but not in the fish, of non-protein-coding DNA sequences of unknown function.

The mouse genome also contains sequences that are repeated many times. This was revealed by the observation that, if we use a randomly cloned 1–2-kb DNA segment as a probe and label it with a fluorescent dye, in most cases this probe will hybridize with several chromosomal regions, indicating extensive redundancies.

Finally, if we consider that there are between 20,000 and 30,000 genes in a mouse genome (which is a reasonable guess) and only 4,377 genes in *E. coli*, this indicates that the average gene density in the mouse is much lower than in the bacterium (~1/100 kb in the mouse versus roughly ~1/1 kb in the bacterium). All these observations support the idea that a large proportion of the mouse genome does not code for proteins and may represent what Susumu Ohno called "*junk*" DNA (Ohno 1972)—unless we find that part of the DNA in question serves other functions that might be important.

Considering all these issues (i.e., a genome with a large size, with a heterogeneous structure, with many redundancies and a large amount of possibly "*junk*" DNA), scientists were then warned from the beginning that unraveling the complete sequence of a mammalian genome would be a long and difficult enterprise.

5.2.2 How Was the Mouse Genome Sequenced?

There are basically two strategies for sequencing a complete mammalian genome. The first one, known as *hierarchical shotgun sequencing* (HSS), makes use of cloned DNA with large inserts such as bacterial artificial chromosomes (BACs—with 150–250 kb DNA inserts), P1 phages or, less frequently, yeast artificial chromosomes (YACs—200–1,000 kb). As explained in Chap. 4, these clones of DNA are assembled into a series of overlapping elements known as *contigs* (from contiguous DNA segments), which altogether make a physical map encompassing chromosomal segments of the greatest possible dimension. The DNA clones mentioned above are generally selected once they have been thoroughly checked for structural integrity, rejecting those that are chimeric or have deletions (a situation that is common in YACs but less common with BACs). The assembly of these cloned DNAs into contigs is achieved by careful fingerprinting of each and every clone. When the contigs are established, in general from several individual clones ranging from 100 to 1,000 kb, a subset of minimally overlapping clones is chosen and each of its elements is sequenced several times to minimize the effect of sequencing errors (this minimal set is sometimes called the "*Golden Tiling Path*" or simply the "*Golden Path*"). The primary sequence is called a *read* and the released genome sequence, or *draft*, results from the integration of several independent reads (in general 10–15, sometimes more). After computerized processing of these independent reads, and if we suppose that the sequencing errors occur randomly, the final rate of errors in a given consensus sequence is very low, in general less than one error per 10^5 bp.

The HSS strategy is relatively slow and tedious, but it is systematic, progressive and highly reliable. The use of clones with large DNA inserts is also a way to circumvent, at least to a certain extent, the difficulties associated with the sequencing of DNA repeats and variations in copy number, which are true nightmares for sequencers. However, the HSS strategy has the disadvantage that only long DNA fragments cloned in a vector can be sequenced. Unfortunately, it is virtually impossible to clone the whole of a mammalian genome in BAC or YAC vectors, for reasons that are associated with both the structure of the DNA in some chromosomal regions and with the cloning technology.

A second strategy, called *whole-genome shotgun* (WGS), consists of the mechanical fragmentation (e.g., by sonication) of the mammalian DNA into segments measuring 100–400 bp, which are sequenced from both ends using the *chain termination method*. Multiple reads of the targeted DNA are obtained by performing several independent rounds of this fragmentation, each followed by sequencing. Once the sequence of the targeted DNA is achieved, computer programs are then used to assemble the pieces of the puzzle, ordering the individual fragments into virtual contigs, then in super- or hypercontigs and finally in ultracontigs based on the overlapping sequences of the different reads.

The WGS method is fast and (in theory) does not require the pre-existence of a physical map. Unfortunately, it does not allow the sequencing of certain genomic

segments such as highly repeated regions. Combining the two strategies (WGS first, then HSS) allows for the correction of most of these difficulties. In short, the two strategies are complementary: WGS provides rapid and relatively good coverage early in a project, while HSS is more systematic and more efficient for the sequencing of regions with repeated sequences. The human genome was sequenced by using mostly the HSS strategy, while the mouse and all other mammalian genomes were sequenced by using mostly the WGS strategy, with the help of HSS only for finishing some regions.

In fact, technical and methodological difficulties emerge when the objective is to sequence the genome of a species for the first time (the human genome in this case), but the situation is greatly simplified when the project is to sequence the genome of evolutionary related species. This is because it is possible to take advantage of the existence of the many interspecific structural homologies that exist at the chromosomal level. Thus, the mouse and rat genomes were sequenced mostly by WGS, and accordingly were completed much faster than the sequencing of the human genome (Fig. 5.1).

Sequencing techniques have progressed enormously recent years and many steps are now fully computerized, reducing human intervention and cost. The latest assembly released by the Mouse Genome Sequencing Consortium (MGSC) has a length of 2,730,871,774 bp (Golden Path from *Ensembl*—September 2013). Curators of the database consider that at least 99 % of the mouse genome sequence is established, with the exception of only a few small gaps (~180) scattered in between a total of 750 contigs, with less than one sequencing error per 10^5 bp. All of the chromosomes have been entirely sequenced, including the X and the Y, allowing comparisons with homologous regions of the human and other mammalian genomes to be performed at a very high resolution.[1]

Such comparisons, revealing similarities and differences, are a rich source of information. Similarities (i.e., sequence conservation), as we will discuss later, allow us to detect regions that are very likely under selective pressure and which, for this reason, have remained unchanged or nearly so for millions of years, indicating that they are presumably genetically important and, accordingly, have resisted random drift. Differences at the sequence level may be even more interesting a priori, because they may contain information explaining how speciation proceeds. It will be obviously interesting to discover both the mechanisms governing these processes and the consequences of these differences at the phenotypic level. We will come back to this point several times, which is well exemplified in the case of variations in gene or DNA copy numbers (copy number variations or CNVs, see Sect. 5.3.6.).

The mouse sequencing project was undertaken by the MGSC, an organization that consisted originally of three laboratories: the Whitehead Institute for Biomedical Research at the Massachusetts Institute of Technology (USA), the Washington University Genome Sequencing Center (USA), and the Wellcome Trust Sanger

[1] The mitochondrial DNA has also been sequenced. See Sect. 5.6.

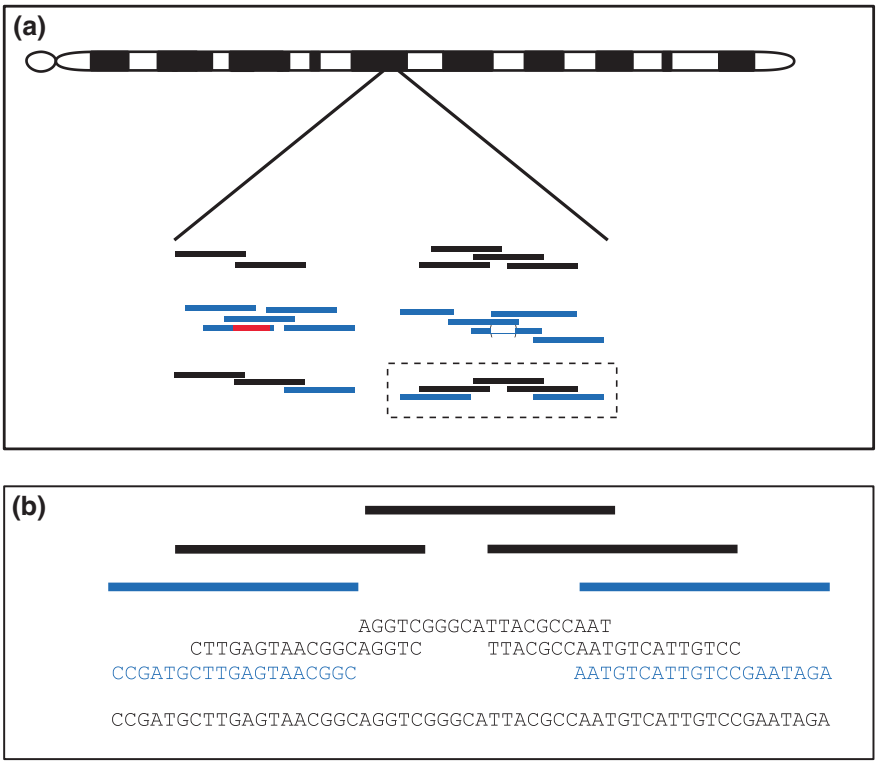

Fig. 5.1 *Strategies used for sequencing mammalian genomes*. Two strategies have been used for sequencing the mammalian genomes: hierarchical shotgun sequencing (HSS) and whole-genome sequencing (WGS). HSS (Fig. 5.1a, b) has been used for sequencing the human genome. It works in two successive steps and makes use of bacterial artificial chromosomes (BACs, ~150–300 kb) or yeast artificial chromosomes (YACs, ~500–2,000 kb) that have been previously used for the establishment of the physical map or "contig map". In the first step (**a**), the integrity and quality of these cloned DNAs is carefully checked (absence of mosaicism, absence of deletion). Then the most interesting elements (**b**) of these contigs (those representing the "*golden path,*" with minimum overlap) are completely sequenced and the sequence ordered. The HSS strategy is systematic and reliable, but it is slow and does not allow the sequencing of regions with repetitive DNA. The whole-genome sequencing strategy (WGS) (Fig. 5.1c, d, e) has been used for sequencing most of the mouse genome. This strategy completely bypasses the BAC/YAC step and consists of the direct mechanical fragmentation of DNA samples to obtain a mixture of independent, randomly cut stretches of DNA 100–400 bp long (**c**). These stretches are then cloned using adaptors, labeled, and sequenced end-to-end (**d**). In a third step (**e**), sequence overlaps are looked for by using appropriate computer software and the clones are then arranged in a head-to-tail manner to form virtual contigs of non-redundant, top-quality sequences. In the final step, the contigs are anchored to the specific chromosome they belong to. The process is generally repeated several times to reduce the number and size of the unsequenced regions and strengthen the quality of the sequence. The gaps in the sequence resulting from the WGS strategy are filled, where possible, by HSS. In the current mouse sequence, the number of gaps is extremely reduced

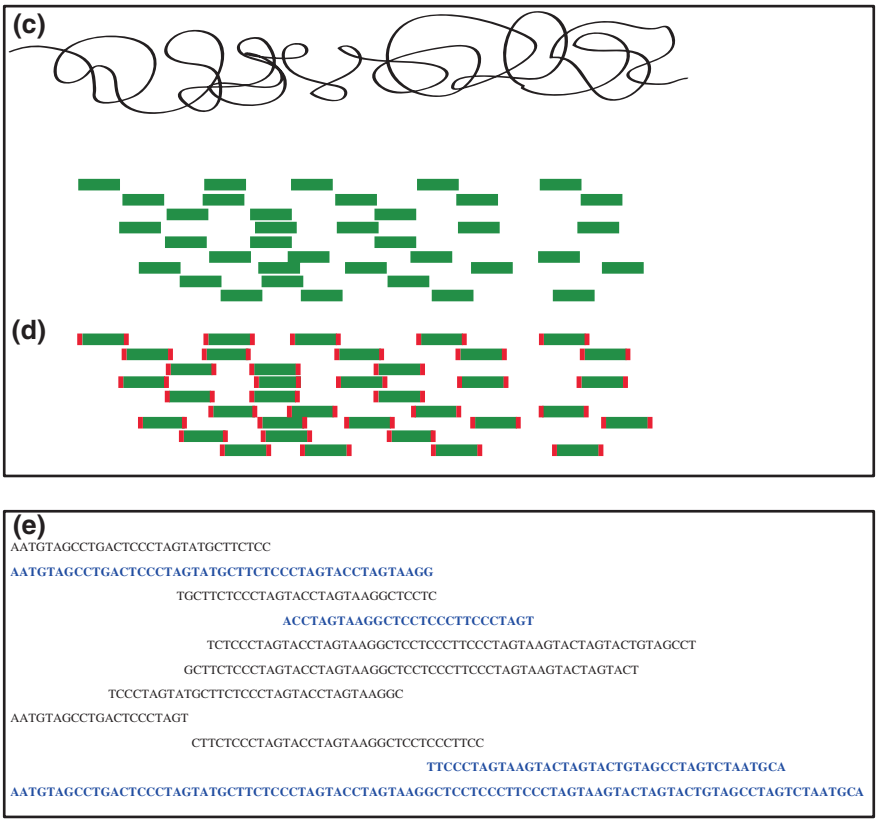

Fig. 5.1 (continued)

Institute (UK). Based on discussions with the scientific community, MGSC investigators decided to sequence, first, the genome of a female from the C57BL/6 inbred strain. At the same time, four other inbred strains (A/J, DBA/2J, 129X1/SvJ, and 129S1/SvImJ) were being sequenced by the CELERA firm in another independent WGS project. Here again, interstrain comparisons have been of great interest when matched with particular phenotypes. Nowadays, the original projects are finished, even though molecular biologists at the MGSC keep working on some specific regions. The Mouse Genomes project from The Wellcome Trust Sanger Institute recently completed the sequencing of an additional 17 inbred mouse strains: 129P2, 129S1/SvImJ, 129S5, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CAST/EiJ, CBA/J, DBA/2J, LP/J, NOD/ShiLtJ, NZO/HiLtJ, PWK/PhJ, SPRETUS/EiJ, and WSB/EiJ (see http://www.sanger.ac.uk/resources/mouse/genomes/). These strains were very carefully selected after extensive discussions via the Internet among the members of the community of mouse geneticists. The genome of the FVB/N inbred strain, popular for the production of transgenic animals and for skin carcinogenesis studies, is now also available (Wong et al. 2012).

These genome sequencing projects are now benefiting from new, ultra-efficient sequencing technologies known as *next-generation sequencing* (NGS). It is likely, for example, that many genome sequences from highly informative strains (strains from the Collaborative Cross, for example—see Chaps. 9 and 10) or even some carefully selected individual mice will become available, contributing efficiently to the analysis of complex traits. Even if the development of bioinformatics resources for the interpretation of the tremendous and ever-increasing amount of data remains a challenge, we can say that the mouse genome-sequencing project is, without any qualification, a complete success from an analytical point of view. However, from now on scientists will have to consider a new challenge, at least as important: the annotation of all sequences in this genome. No doubt they will be kept very busy for another few years.

5.3 The Structure of the Mouse Genome

Once a genome is entirely sequenced and the sequence stored in a database, scientists can then start looking at it in more depth. This structural analysis, run in parallel with a functional analysis, is part of the so-called *genome annotation process*, and one of the first challenges in this matter is to identify and characterize as accurately as possible the DNA regions containing the genes proper (i.e., the DNA coding for proteins or RNAs), the regulatory elements, and some other potentially important structures. This is a real challenge because, if we recall what we said earlier when discussing gene density in mammalian genomes, the protein-coding and related sequences represent only a very small proportion of the mammalian DNA. However, if we consider that this functionally important fraction of mouse DNA, because it is under the constraint of *purifying* (i.e., negative) *selection* during evolution, is likely to be highly preserved across different species, we already have outlined a strategy to identify and estimate it. This estimation has been achieved, shortly after the release of the first draft of the mouse sequence, by cross-comparing several regions of the human genome with various short sequences of the mouse genome, and the answer was that there is indeed great interspecific homology (over 95 %) for around 3.5–5 % of the genomic DNA sequences. There are good reasons to believe that the genes encoding proteins and other important sequences are gathered in this fraction.

5.3.1 Finding the Coding and Related Sequences

The first step in the process of genome annotation generally consists of checking for the presence or absence in the newly sequenced genome of some specific sequences previously characterized in other species (the exons, for example), and

evaluating the number of copies, their organization and flanking sequences, etc. The geneticist may also wish to make an inventory of all the genes of a given species: those encoding proteins and those transcribed only into RNAs. These questions have triggered a multitude of intensive studies, many of which have now resulted in more or less precise answers.

5.3.1.1 Retrieving Specific Sequences

Nowadays, finding a particular sequence in a genome is relatively easy and several software packages have been designed for this purpose. The most popular is BLAST. BLAST allows similarity searches to be performed against any databases of recently sequenced organisms. BLAST will rapidly identify and retrieve a sequence in the human or rat genome that resembles a mouse sequence based on similarity of sequence. These software packages work, roughly, like the sub-programs that are activated when, working on a text file, one selects the command "*Find*" to specifically retrieve a chain of characters, with the important difference that BLAST can retrieve sequences that are not 100 % identical to the queried one. ROSETTA[2] and SEQUENCHER® sequence analysis softwares are other packages useful for finding genes (and not only coding sequences) by comparisons, for example, between human and mouse DNAs. ROSETTA performs sequence alignments and compares the exon sizes, splicing sites, etc., and finally makes gene predictions.[3]

When a coding sequence (a mouse exon, for example) is used as a template for retrieving the most closely related sequences in the human or rat genomic sequence, in ~95 % of the cases BLAST retrieves a sequence with high similarity and 90 % of these sequences are on the homologous chromosomal segment in all three species. Geneticists say that they share the same *syntenic* location (from the Greek, meaning "on the same ribbon") and these genes are called 1:1 *orthologs*. This indicates that most of the genes in a given mammalian genome are part of an ancestral heritage and do not vary much among other mammalian species even if, sometimes, there are variations in terms of copy numbers, as we will discuss further. Differences in terms of presence versus absence are rare but occasionally occur. For example, approximately one hundred predicted mouse genes identified in the initial mouse draft sequence were reported as missing (having no homologous counterpart) in the human genome. The reverse of course is also true, and some human genes are absent in both the mouse and rat genomes. A good example of such a situation is the gene encoding human interleukin 8 (*IL8*), which cannot be found in the rat and mouse regions of homology for HSA-Chr 4 (see Fig. 5.2).

---

[2] https://www.rosettacommons.org/.

[3] Sequencher version 5.1 sequence analysis software, Gene Codes Corporation, Ann Arbor, MI USA http://www.genecodes.com.

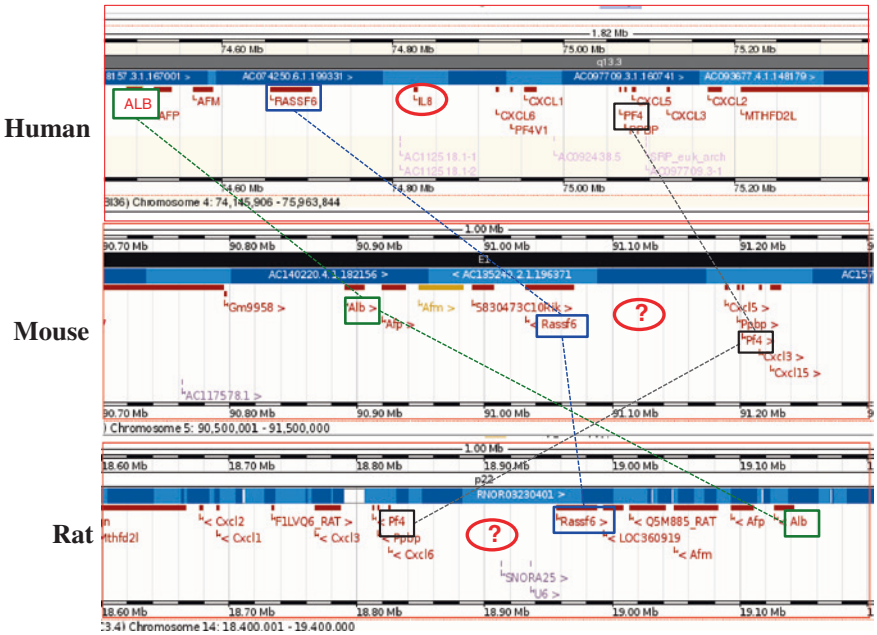

Fig. 5.2 *Sequence comparisons between mammalian genomes*. The orthologous copy of the human gene encoding interleukin-8 (*Il8*) is missing in the mouse and rat genomes. The figure shows the region of human chromosome 4 (HSA4) where the *IL8* gene is located, with the homologous regions in mouse chromosome 5 (MMU5) and rat chromosome 14 (RNO14). The rat chromosome is affected by a paracentric inversion when compared with the human and mouse homologous regions. Such rearrangements are extremely common in the mammalian genomes and are very useful (with other methods) for establishing the phylogenetic relationships between species. The images are from the *Ensembl* Genome Browser database (May 2013)

These qualitative differences are not easy to explain and can result either from true deletions, with no consequences at the phenotypic level, or from the fact that the supposedly deleted genes in fact still exist elsewhere in the genome but have evolved so rapidly, in one or the other lineage, that they are no longer recognizable as orthologs based on sequence comparisons. The first hypothesis is the most likely, since such segmental deletions of recent origin have been discovered, by chance, in the genome of several inbred strains while others were reported normal (undeleted). For example, mice of the C57BL/6JOlaHsd substrain (also known as C57BL/6S) are homozygous for a deletion encompassing the entire α-synuclein gene (*Snca*-Chr 6) (Specht and Schoepfer 2001). These mice are fertile and have a normal lifespan, but they have at least one gene inactivated compared with most other C57BL/6 substrains. Examples of this kind have been reported in many other laboratory inbred strains and also exist in the human and rat species (Perez et al. 2013).

Finding genes in the genome of one species, once the orthologous versions of these genes are known and already identified in the genome of another closely

related species (such as human, rat, and mouse), is then relatively straightforward and many computer programs can do this, even if surprises and difficulties occasionally occur, as we will see later.

5.3.1.2 Identification of the Coding Sequences

The situation is more complicated when the objective is to identify all the coding sequences (all the exons, for example) in a freshly sequenced genome.

A first and relatively efficient technique, known as *exon trapping*, was published in 1991 (Buckler et al. 1991). With this technique, a cloned genomic DNA was inserted, by genetic engineering, into an intron of the human immunodeficiency virus 1 (HIV-1) *tat* gene (Trans-Activator of Transcription), contained within the plasmid pSPL1. COS-7 cells were then transfected with these constructs, and the resulting RNA transcripts were processed in vivo. The splice sites of exons contained within the inserted genomic fragment were put in phase with the splice sites of the flanking *tat* intron. The mature RNA collected from the COS-7 cells contained the potential exons, which could then be amplified via RNA-based PCR and ultimately cloned.

Exon trapping has been a very helpful technique, especially in the projects whose aim was the positional cloning of a gene identified only by a mutant allele. However, compared to more recent techniques, it has two major drawbacks: (*i*) it does not trap faithfully the large or very small exons, and (*ii*) it is relatively expensive because it requires a significant amount of bench work and in vitro cell culture.

5.3.1.3 Using Expressed Sequence Tags (ESTs) for the Detection of Transcribed Sequences

Taking into account the fact that several mammalian genomes are now entirely sequenced, strategies have been developed that are based on the identification at the genome level, by all possible techniques, of sequences deduced from transcribed products. One of the first strategies consists in using so-called *Expressed Sequence Tags* (ESTs). ESTs are short sub-sequences of cDNA corresponding to a few hundred (~350–500) base-pairs of a cDNA, starting from the 3' end, sometimes from the 5' end. Millions of such ESTs (from several mammalian species) are available in public databases, and the sequence of each of these ESTs can be used as a molecular probe to retrieve the complete sequence of the gene the EST belongs to (or is related to), simply by "pulling on" the flanking sequences. Since the ESTs stored in a given database were in general prepared from a specific tissue (brain, blood, skin, neoplastic tissue, etc.) at a certain step of development (embryonic, 10 days, adult, senescent, etc.), using these ESTs for gene identification has the additional advantage of providing information concerning the transcriptional level and the gene expression pattern for the annotation process. ESTs have been

instrumental for the initial identification of many genes in the mouse as well as in the human genome, and still are. In addition, the sequence alignments can be performed entirely in silico, which means rapidly and at virtually no cost. The major drawback of these ESTs is that only a fraction of the genes are expressed simultaneously, and consequently the EST collection in a particular database represents only a fraction of the genes of a given species. Finally, some genes are transcribed only in particular circumstances, at very low levels, or transiently and, by definition, they are poorly represented in EST libraries or databases.

5.3.1.4 Using Strategies Based on Artificial Intelligence

Other strategies, requiring sophisticated informatics, rely on the identification of some transcription-related motifs that are part of most protein-coding genes (Blanco and Guigo 2005; Harrow et al. 2009) (see also next section). These motifs have been successfully used for gene detection with software systems like GENSCAN, developed by Burge and Karlin (1997). In addition to the strategies mentioned above, more refined prediction programs, often referred to as de novo or ab initio *gene finders*, have also been developed by geneticists and computer scientists. These programs are based on the existence of subtle differences at the sequence level that can be used to sort out putative coding regions from non-coding regions by making use of the so-called hidden Markov chain models. These prediction models are based on the fact that biases and dependences exist in coding sequences that are not observed in non-coding regions. This means, for example, that the five preceding bases influence the probability of finding a particular base at the sixth position of a new sequence if, and only if, the sequence in question is a coding sequence. When scanning a novel nucleotide sequence, the program computes a *coding likelihood score*, based on a Markov chain model of order 5, and makes an assessment as to whether the sequence is more likely to be from an intron, exon or intergenic region (Harrow et al. 2009).

All these sequence prediction algorithms are being constantly improved based on the experience acquired from training with DNA samples whose sequence is fully annotated. These programs work more or less like the software designed for language translation. Years ago, the meaning of "computer-translated" sentences was only remotely related to the meaning in the original sentence and sometimes limited to an unordered set of key words. Nowadays, the quality of the translation is very good (at least for certain languages). Based on their encouraging results, researchers consider that, as of today, around 85 % of genes should be rapidly and easily detected in any new mammalian genomic sequence by using software of this kind. Most of these newly discovered genes must, however, be validated by other approaches because the discovery of a gene-like structure does not automatically mean that an authentic, indisputable, and functional protein- or RNA-encoding gene has been "fished". This validation is very important work, whose aim is to create a gold-standard reference for gene annotation. A program of this

kind has been undertaken by the Human and Vertebrate Analysis and Annotation (HAVANA) team at the Sanger Institute, where the human, mouse, and zebrafish genomes are carefully annotated manually.

Making sequence comparisons (or alignments) with other genomes (human, rat, zebrafish) has allowed a rather rapid identification of a great number of mouse genes. However, from now on, the identification of novel genes in the mouse will probably progress at a somewhat slower pace because the situations researchers face are sometimes difficult. Some genes, for example, are very large and extensively fragmented, while others are very small with only one intron or even no intron at all (for example, the intronless genes encoding RNAs and histones). Since neither of these two categories of genes correspond to the "canonical" representation of most mammalian genes, they have to be annotated manually and this takes much more time. Another very common situation is that, although they share a syntenic location as expected, orthologous genes are not always in a 1:1 ratio but rather in 1:2, 1:3, and so on. We will describe situations of this kind, where the "pseudo-orthologous" copies are sometimes slightly altered or incomplete, but are still transcribed and accordingly annotated as a true gene.

Finally, overlapping and nested genes have been shown to exist in mammals just like in *Drosophila,* with various imbrications of their structure with their neighboring genes. Nested genes were generally described as genes with a relatively short size, consisting in general of only one exon and entirely nested within a single intron of a host gene. The situation has recently changed dramatically as a consequence of more in-depth analysis of the mouse transcriptome, as we will discuss further in this chapter, and many RNAs are transcribed from the mouse genome whose function is not yet established. In the same way, genes have been found that are transcribed in the opposite orientation to their neighboring host genes, and sometimes negatively influence the transcription of these genes via antisense-mediated inhibition (see below—Chap. 6 on X-inactivation). Identification of nested genes is difficult but, fortunately, approximately 60 % of nested genes are conserved in mouse and human in the same genomic context.

The ENCODE project (ENCyclopedia Of DNA Elements), which is essentially the next step for the Human Genome Project, has set as its major aim the establishment of all the structural and functional elements of the genome. It is definitely an ambitious project but it makes a lot of sense and is really necessary if we consider its potential applications. Here again, just like for the sequencing of the mouse genome, we can say that this meticulous analysis conducted at the DNA level would have to be achieved one day because the general feeling of the community is that it is a crucial endeavor, if not simply the essence of genetics: then why not do it right now, as rapidly as possible, on a systematic basis?

The preliminary results of the ENCODE project, although still fragmentary, have already changed our understanding of the mammalian genome by demonstrating that the mammalian DNA hitherto labeled "*junk*" might not be *junk* after all.

5.3.2 The Canonical Architecture of a Protein-Coding Gene

As discussed in the preceding section, many points remain to be elucidated concerning the structural organization of the mouse genome. However, as of today, hundreds of genes have been entirely sequenced in several species including mouse, rat, human, and domestic animals. As a result, it is now possible to outline the classical or canonical architecture of the "average" mammalian gene.

A gene is a segment of DNA that encodes an RNA molecule that may or may not be translated into a protein. For this reason, geneticists formally distinguish two types of genes: the protein-coding genes and the non-protein-coding genes. For many years, and up to relatively recently, molecular geneticists considered that the two strands of the DNA molecule were not equivalent: one of them was the coding strand while the other was the template or anticoding strand. However, and unexpectedly, it has recently been demonstrated that mammalian DNA is pervasively transcribed from both strands. We will come back to this important point later in this chapter when discussing the transcriptome and the non-protein-coding RNAs. Here, we will simply discuss the organization of a classical protein-coding gene as it has been established as the result of thirty years of careful positional cloning, sequencing, and annotation.

The transcription of a protein-coding gene into a primary mRNA proceeds from the 5' to the 3' end and starts ~50–60 bp upstream of the first AUG codon, encoding a Methionine. The ~50 bp between the transcription initiation site and the initial AUG is part of the so-called 5'-untranslated region (5'-UTR) or *leader sequence.* This sequence usually contains a ribosome binding site (RBS), known as the *Kozak sequence* (gcc)gcc(A/G)cc**AUG**G (Kozak 1987), that includes the AUG initiation codon.

Upstream of the transcription start site at the 5' end, several consensus sequences have been identified that are part of the promoter sequence of the gene, as we will see in the next section of this chapter. Opposite to the 5'-UTR is the 3'-UTR or *trailer sequence,* required for the processing of mRNA, the size and canonical sequence of which is not as precisely known as that of the leader sequence. The end of a structural gene is called the *transcription termination site*. Some specific sequences are also found in the 3'-UTR. First is a *polyadenylation signal* composed of sequences like AAUAAA or a slight variant. The polyadenylation signal indicates that transcription will be terminated approximately 30 base-pairs downstream of it, while a tail composed of a few hundred adenine residues (the poly-A tail) will be added to the transcript. The poly-A tail is important for the nuclear export, translation, and stability of the mRNA.

Since 1977, it has been established that many (around 60 %) mammalian genes have a heterogeneous structure: some parts are included in the final protein or RNA product, while others have another destination or are merely degraded. Hence, the coding sequences of most mammalian genes are composed of an alternation of exons (*ex*pressed regi*on*) and introns (*intr*agenic regi*on*). Introns are spliced off during RNA processing or maturation when the pre-mRNA becomes

a mature mRNA, ready to be translated into a protein product. RNA splicing is a complex and very precise procedure that is regulated and controlled at the cellular level, at the base-pair level of precision (Fig. 5.3). This process requires several highly specific tools: at least five small nuclear RNAs and around 150 proteins, collectively known as the *spliceosome* (Hoskins and Moore 2012). Among the most important are the small nuclear ribonucleic acids or snRNAs, the small nucleolar RNA or snoRNAs, and specific enzymes including the ribonucleoproteins or "snurps".

Splicing sites can be identified at the DNA level because they have a consensus sequence: the first two bases at the beginning of an intron (at the 5' end) are almost always GT and the last two, at the 3' terminus of the same intron, are almost always AG. The sequences immediately upstream of the AG and downstream of the GT are also conserved, although to a lesser degree. For example, the intronic region upstream from the AG is usually a region rich in C and T. The regions at the 5' boundaries of the introns are called the *donor sites* and those at the 3' end are called the *acceptor sites*. We already mentioned earlier that these splicing sites have been used for the identification of exons with the exon trapping technique. Most sequence identification software can also identify these sites in a mouse DNA sequence and label them as "candidate splicing sites" (Fig. 5.4).

Not all exons in a gene are spliced and subsequently assembled to form the final RNA product. In fact, if we consider that the exons correspond to "functional units" and the introns are "spacers" between these functional units, we observe that the exons can be assembled into different combinations to produce different polypeptides. This is known as *alternative splicing* and it is estimated that ~95 % of multi-exonic genes are alternatively spliced in mammalian genomes. From numerous observations, it is also known that the exons in a gene are of two types: (i) those that are always present in all transcripts, which are often referred to as *constitutive or major forms of transcripts;* and (ii) those that are *optional* or *alternative*. Exons of the second type, those that are only included in some spliced

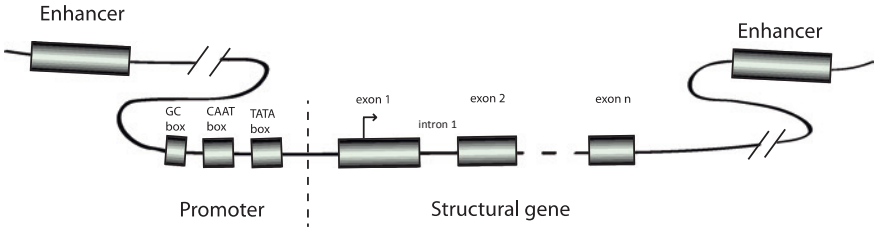

Fig. 5.3 *Canonical (and simplified) representation of a protein-coding mammalian gene*. The enhancers represented in the figure are not always present and are sometimes distant from the promoter region by several Mb. Many sequences in the promoter region are important for gene regulation, but not all of them have been identified and they probably vary from one gene to another. Not all genes have a CAAT box or a TATA box. Finally, not all genes have intronic sequences, and not all exons are represented in the final product

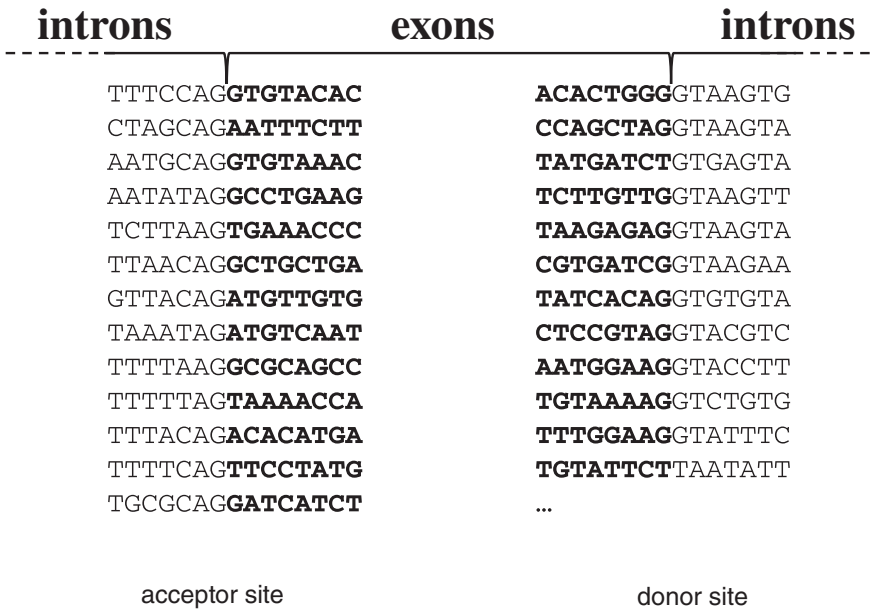

Fig. 5.4 *Splicing sites.* The figure represents the splicing sites found in the sequence of the gene encoding the mouse leptin receptor (*Lepr*-Chr 4). The intronic sequence of the donor (GT) and acceptor (AG) sites are highly preserved (this is known as the GT-AG rule)

forms, as opposed to the major transcript forms, are mostly not conserved across species and are probably of recent origin (Modrek and Lee 2003) (Fig. 5.5).

Alternative splicing can generate a large variety of proteins from the same DNA coding sequence by modifying the exonic contribution of the mature messenger RNAs. It is clear that the actual number of genes in a species has only a relative meaning, since the splicing machinery can tremendously increase genetic diversity. In this context, the number of exons is certainly much more informative for researchers than the number of genes. Alternative splicing is considered to be a very important mechanism for resolving the discrepancy between actual gene number and organismal complexity.

The mechanisms regulating alternative splicing and leading to the incorporation, or lack thereof, of a specific exon into the final product (sometimes designated the *splicing code*) are not yet completely unraveled. These processes probably involve *trans*-acting proteins (repressors and activators) encoded elsewhere in the genome that pair with *cis*-acting regulatory targets on the pre-mRNA. It is also likely that the secondary structure of the pre-mRNA transcripts plays a role in the regulation of splicing (Barash et al. 2010).

As we will discuss in Chap. 7 and as demonstrated by hundreds of positional cloning experiments performed in the mouse and rat, splicing sites are common targets for the occurrence of mutations. These sites are not always in frame with the sub-modulation of the mRNAs in triplet, and can also occur within codons.

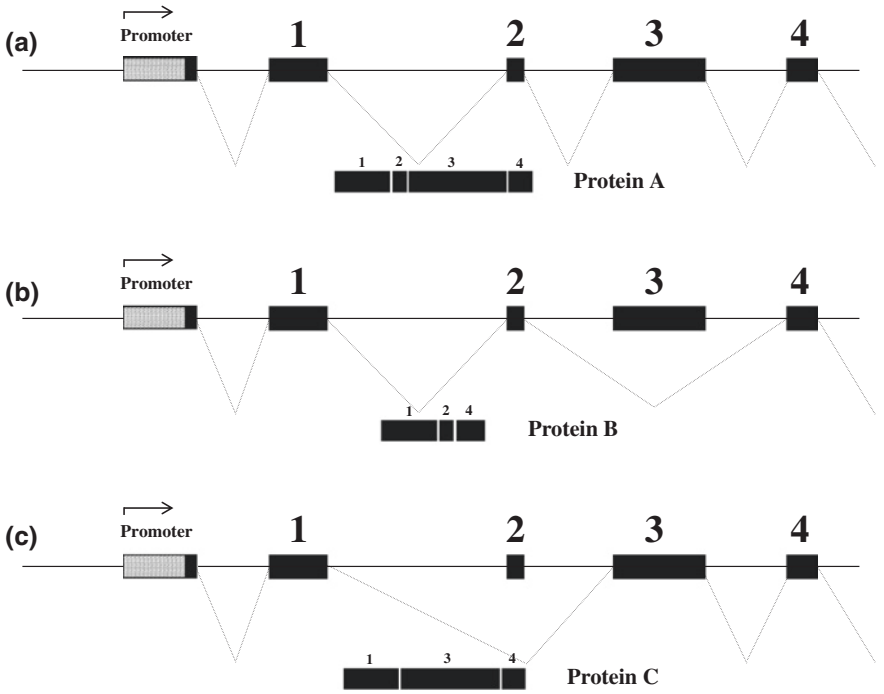

Fig. 5.5 *Alternative splicing.* In mammals, around 95 % of multi-exonic genes are alternatively spliced to produce different proteins (**A**, **B**, **C** …). Some exons are present in all transcripts (*constitutive* or *major forms* of transcripts), while others are *optional* or *alternative*. Exons of the second type are mostly not conserved across species and are probably of recent origin. For orthologous genes, the number of exons is sometimes variable among species, and the presence of recently captured exons is sometimes observed

In these cases, of course, the two contiguous exons are inseparable and are jointly incorporated into the transcript or skipped. The mRNA transcript, once adequately spliced, receives a cap of a methylated guanine nucleotide that is added to its 5' end to protect it.

The enormous amount of information collected by mouse geneticists indicates that the average size of a mouse gene is approximately 30–40 kb at the DNA level, while the average mature or processed mRNA molecule (mRNA mature transcript) is approximately 2 kb. The average gene density is in the range of 1 gene per 95 kb of DNA, i.e., very close to the predictions. The smallest (known) gene is 0.1 kb and encodes the t-RNATyr. The largest gene is *Titin* (*Ttn*-Chr 2), with 2.8 Mb of genomic sequence and 363 exons producing a spliced mRNA larger than 100 kb. The introns are also of various sizes, ranging from around 0.5 kb for the short ones to 30 kb for the longest (dystrophin-*Dmd*), with an average intron size of 4.7 kb. For the exons, the shortest consists of only 9 bp (exon 29 of *Myo5a*), and the largest is 7.6 kb long (exon 26 of *Apob*), with an average exon size of approximately 290 bp. Altogether, when added up, the exons represent

1.2 % of the total mouse DNA, the introns 26.7 %, and the intergenic regions 69.3 %. The number of exons per multi-exon gene varies from 1 to 363 with an average of 8.4. Finally, around 4,000 genes have only one exon.

The configuration of the "typical" mouse structural gene, as we just outlined it, is probably very similar to the average mammalian gene, and this is a blessing for the establishment of comparative maps; in short, the DNA sequence of two (not only one) mammalian genomes is an invaluable tool for making predictions about a third one. Many examples could be obtained from the cross-comparisons of mouse, rat, and human sequences. As of today, 17,054 mouse genes have an orthologous copy in the human genome, while 18,458 mouse genes have an orthologous copy in the rat. Finally, a total of 20,388 mouse genes have orthology annotations with at least one other species.

The classical gene we just described corresponds to a protein-coding gene. In fact, we now know that this category of genes represents only a proportion of the genes in the mouse genome that specialists consider to be in the range of 25–30 %. Most other genes encode RNA molecules of various sizes: some have an *open reading frame* (ORF) but most do not. Some are spliced, others are not, and the majority of these transcripts are processed further in smaller molecules. Most of the RNAs stay in the nucleus, suggesting that they have a function. Finally, all these RNAs exhibit a rather low degree of interspecific homology, indicating that the selection pressure they experience is of a different type. We will discuss this point more extensively at the end of this chapter when discussing the mouse transcriptome.

In November 2014, the Mouse Genome Informatics database estimated the number of mouse genes with nucleotide sequence data at 34,628 and the number of genes with protein sequence data at 24,553. This information seems reliable when compared with other species. Out of these genes, only 16,345 have experimentally based functional annotation.

Finally, we must point out that the distribution of genes in the mouse genome is very uneven. Mouse chromosome 11, for example, has twice the gene density of chromosomes 10 or 12, and the Y chromosome has only a few genes in an "ocean" of repeated DNA.

5.3.3 Finding the Regulatory Sequences

One of the biggest challenges of genome annotation is to identify gene regulatory regions. These comprise proximal and distal regulatory elements, according to their distance from the transcription starting point. Proximally are the *promoters* and associated promoter elements. Distal elements are enhancers, silencers, insulators and locus control regions (Fig. 5.3). Proximal and distal elements are usually composed of clusters of short intermingled transcription factor-binding DNA motifs referred to as modules or cis-regulatory modules (CRMs) (Hardison and Taylor 2012).

DNA sequence and local chromatin landscape act jointly to determine transcription factor (TF) binding intensity profiles. As a result, a regulatory module is defined by its sequence, since it binds transcription factors, and is thus expected to contain specific binding sites for these. It is further defined by its accessibility to TFs, which is linked to chromatin structural specificities such as histone modifications and local occupancy by nucleosomes. These are highly dynamic events which reflect the history of the cell and which are responsible for differential gene expression in animal development and cell differentiation. This implies that canonical binding sites for transcription factors are seldom sufficient to define a regulatory module, and methods relying on binding site identification usually have a high rate of false positives. For example, out of 132 regulatory modules predicted by algorithm analysis to bind TCF4 (a key transcription factor in the WNT1 signaling pathway), only 10 were validated using chromatin immunoprecipitation (ChIP)—little more than a random representation (Hatzis et al. 2008). This further implies that most CRMs will be difficult to identify until the chromatin landscape around them is defined. As a result, whereas the transcriptional apparatus reads the regulatory elements in the genome very efficiently, we still lack a universal syntax to decipher them, and this is quite critical: for regions that are defined by an unequivocal syntax, such as the coding exons, mutations can be characterized by just sequencing the whole mutated genome, together with low-resolution meiotic mapping, using no more than two dozen F2 mice (Xia et al. 2010; Arnold et al. 2011). Reaching the same level of power for regulatory regions would change the face of gene regulation analysis. Fortunately, this field is developing at a rapid pace, following the systematic reliance on strategies that directly measure sequence occupancy by Transcriptional Regulatory Factors (TRFs) within the living cell, such as chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) or DNase I digital genomic footprinting, which are currently performed or compiled by ENCODE (the ENCODE Project Consortium 2011—see above). Most of these results to date have been obtained for human but major conclusions also apply to the mouse, as demonstrated by results already obtained in this species.

Proximal regulatory modules (PRMs) at and around transcriptional start sites (TSSs) are the most straightforward regions to identify, since the TSS is accessible from the transcription product, the RNA. Cap-analysis of gene expression (CAGE) and RNA sequencing (RNA-seq) have contributed to the definition of TSS and consequently of PRMs. From these analyses, it appears that mammalian promoters can be separated into two classes: evolutionarily conserved promoters bearing a TATA box, and more plastic, evolvable CpG-rich promoters. The latter are by far the most frequent promoters since the TATA box (with a core DNA sequence 5'-TATAAA-3') is found in only one quarter of all promoters in a mammalian cell, usually around 30 bp upstream of the transcription start site. The TATA box, the first core promoter element identified in eukaryotic protein-coding genes (Goldberg 1979), is an anchoring site for the pre-initiation complex of transcription involving RNA polymerase II. The CpG sequence works similarly via the Sp1 factor. A CAT (or CCAAT, or CAAT) box, with a

consensus sequence GGCCAATCT, is inserted upstream of the TATA box, 75–80 bp from the transcription start site. Some genes with relatively ubiquitous expression do not have this GGCCAATCT sequence. In CpG-rich promoters, the start sites are usually multiple and organized in clusters at the 5' end of the gene, whereas TATA-box-bearing promoters have a single or at least a predominant start site. As of 2006, 729,504 potential mouse TSS sites were defined, organized in 159,075 clusters, a figure that far exceeds the number of genes identified (see above) (Carninci et al. 2006). Furthermore, mapping techniques such as CAGE are quantitative and provide a measure of the amount of transcription initiation in any given genomic region or for a given gene, in different tissues.

The situation is much less clear for distal regulatory elements such as silencers, enhancers or locus control regions. Enhancer elements, which can be located at some distance from the core promoter elements, where the transcription initiation apparatus is bound, are sites for fixation of transcription factors. The enhancer-bound transcription factors bind co-activators such as Mediator and p300, which in turn bind the transcription initiation apparatus, thus providing a link between enhancers and promoters. Non-coding RNAs may be associated with Mediator in this process (Lai et al. 2013).

Constraints on distal regulatory elements appear rather loose. Enhancers have been located at the 5' or 3' ends of coding regions, within introns and even within coding exons (Birnbaum et al. 2012), where they impose a further layer of constraint on the coding sequence. They can be close to the transcription start site or, in contrast, extremely remote (one to several megabases)—not to mention the possibility of them lying on a separate chromosome, from which they act in *trans* on a gene-coding region (Savarese and Grosschedl 2006). Furthermore, there is no evidence that the closest enhancer to a gene is the one likely to be active on this gene (Li et al. 2012). In cases when regulatory modules are remote, mutations that affect them may lie within another gene. For example, the CRM driving *Sonic hedgehog* (*Shh*) expression in the limb lies within the intron of another gene, *Lmbr1*, which for some time puzzled geneticists (Hill 2007). Similarly, a *Gremlin1* (*Grem1*) CRM lies within the *Formin* (*Fmn1*) gene, such that the latter was long considered as responsible for a limb defect (its original name was *Limb deformity*), whereas it does not have any known function in limb development, contrary to *Gremlin*.

These difficulties will be overcome when most regulatory regions have been defined according to transcription factor occupancy using strategies such as ChIP-seq. There is still a long way to go: according to experiment matrices recently published by UCSC Genome Bioinformatics, only 13 out of about 60 known histone modifications and 120 out of the estimated 1,700–1,900 transcription factors have been examined to date in the human genome by ChIP-seq. These, furthermore, have been analyzed in a number of cell lines in culture (which often bear little similarity to cells within organisms) or in readily accessible adult cells, such as blood cells, but many tissues in the adult, not to mention embryonic stages, have not been investigated—and the mouse genome lags

behind. Tissues, especially embryonic tissues, provide only sparse material, and methods will have to be miniaturized before they can be extensively analyzed. Nevertheless, the power of these new strategies is such that we can be confident that, in the near future, the regulatory syntax of the genome will be worked out. One major difficulty that may remain is attributing a given CRM to a specific gene in a defined physiological or developmental context, since, as we have seen, enhancers may be very remote and there is evidence that the closest enhancer to a gene is not necessarily active on that gene. Assessing the correlation of the chromatin state at enhancers and RNA-PolII occupancy at promoters, for each possible enhancer–promoter pair of elements in a chromosomal domain, may help define enhancer–promoter organization (Shen et al. 2012). This may be insufficient, due to the properties of enhancers discussed above. We see that the regulatory sequences of a gene can hardly be circumscribed a priori. At this stage, genetic approaches may prove very helpful, since, following mutagenesis, a phenotype attests to an alteration that affects one gene with no a priori hypothesis on the regulatory mechanisms for this gene. Unfortunately, ENU mutagenesis is much more efficient at mutating coding sequences than regulatory sites, for reasons that are not entirely clear. It may be because regulatory regions are often redundant (Lagha et al. 2012), and there may be multiple TRF binding sites within an enhancer, making it unlikely that a single mutagenesis experiment will abolish all the binding sites. In contrast, exceptions to this rule have proven highly educational. This is the case for the limb-specific regulatory module of *Shh*, which is located nearly 1 Mb (~0.6 cM) upstream of the coding region and has been extensively characterized via genetic strategies.

These strategies take advantage of several assets of genetic tools. First, they allow a fine mapping of the genetic alteration. This may be very valuable in the case of distant regulatory sequences. It should nonetheless be kept in mind that CRMs are often too remote for molecular walking strategies along the chromosome, but too close for genetic segregation and localization. While a huge number of polymorphisms have been defined in the mouse genome (SNPs), the precision of mapping still depends on the possibility of getting them to segregate in a cross—i.e., the number of meioses that can be analyzed (with 1,000 meioses yielding a 0.1 cM precision). In a historical attempt to localize *Hx*, a limb mutation that turned out to affect the distant CRM of *Shh*, analysis of more than 2,000 meioses in a cross involving *Mus m. castaneus* reduced the candidate region to a little more than 400,000 bp—a genetic tour-de-force, but still insufficient to identify a causative point mutation. At a minimum, genetic mapping based on segregation defines boundaries within which the regulatory sequence can be sought by other approaches.

To characterize the affected sequence in a mutant, an essential strategy is the reliance on multiple alleles for the mutation. It is even better to have alleles of a different nature in addition to point mutations (insertions, translocations), to allow easier entry points into the mouse genome. Thus, for the *Shh* CRM, as for *Gremlin 1* (*Limb deformity—Grem1*^{ld}, another limb mutation), a transgene

insertion provided an entry point to the CRM (Lettice et al. 2002). This illustrates the value of mutagenesis strategies that generate chromosomal accidents (deletions, translocations, transposon insertions—see Chap. 3) to locate regulatory modules. Examples include *PiggyBac*, *Sleeping beauty* or *Tol2* transposons, and ethyl methane sulfonate (EMS)-induced deletions in ES cells (Munroe and Schimenti 2009).

It has been shown over the past few years that many CRMs are active on more than one gene, defining so-called "regulatory landscapes". Thus, many genes in the landscape show the same expression profile as the gene of interest and may be suspected to encode proteins acting in *trans* as regulatory factors. Examples such as *Shh* (CRM within the *Lmbr1* gene) and *Grem1* (CRM within the *Fmn1* gene) are illustrative in this respect. In such cases, it is essential to define whether regulation occurs in *cis* or *trans*, and, up to very recently, only genetic tests could unambiguously settle the issue. The principle of the test is straightforward, but requires that two allelic forms of the regulatory region and its target, respectively, can be discriminated in a genetic cross. When the regulatory sequence is defined by a mutation, this provides the differential allele for the CRM. The gene acted on must have two alleles, either coming from different mouse subspecies or one being an engineered allele. The ultimate demonstration that the characterized alteration in the genome is the cause for the abnormal phenotype will be provided by recapitulating this phenotype using the altered sequence in a functional test, such as expression of a reporter in a transgenesis experiment, or phenotypic rescue by BAC transgenesis, or de novo creation of the suspected mutation by homologous recombination.

With its very powerful tools (different mutagenesis strategies to generate different types of mutations, screens to identify new dominant and recessive mutations, *cis–trans* tests, etc.), genetics could play a major role in the identification of new regulatory modules. However, genetics now has strong competitors over the whole spectrum: targeted mutations, long-range haplotyping by genome sequencing strategies, and identification of remote regulatory modules by scanning the genome via overlapping transgenes. Even before we can directly identify CRMs using appropriate algorithms, genetic approaches may be outdated by genomic strategies—which also are considerably less expensive.

5.3.4 Organization of Syntenic Regions at the Chromosome Level

As we explained in the previous chapters (Chap. 4 in particular), the linear arrangement of mouse genes along the chromosomes tends to be preserved, at least to some extent, among the different species of mammals, recalling the existence of a more or less distantly related common ancestor. This means that when two genes are found closely linked in the mouse, they have a good chance of also being linked in the rat and in human genomes, depending on the degree

of linkage. With the ever-increasing resolution of genetic maps and the availability of genomic sequences of several different mammalian species, it has become possible to reconstruct the progressive reshuffling of the chromosomal segments that occurred across the species in question during evolution. For example, scanning the human, mouse, and rat genomes at high resolution we find that there are 280 orthologous chromosomal segments between human and mouse, 278 between human and rat and 105 between rat and mouse. Comparisons between dog, cat, and cow, whose genomes are also completely sequenced, indicate that the number of chromosome breaks between human and rodents (~280) is consistent with the number of synteny breaks observed in other species separated by similar evolutionary distances. However, the number of chromosomal rearrangements between rat and mouse seems to be excessive if the divergence between the two species really occurred 12–14 Myr ago. Explanations for this discrepancy are lacking.

The existence of these homologies of synteny indicates that, during evolution, many genomic segments of the different species have been broken and then translocated, inverted, or transposed several times. This, however, is difficult to reconcile with the experimental observations presented earlier, indicating that most alterations in the karyotype structure are in general strongly counterselected by impeding normal gametogenesis in heterozygotes. Here again, explanations are awaited to reconcile all these observations, but it is tempting to speculate that this may be linked to the mechanisms of speciation themselves.

Homologous chromosomal segments display great variations in size across the different species. Mouse chromosome 11, for example, contains a large homologous region (almost all) to human chromosome 17q, while some other homologous chromosomal regions are extremely small-sized, and are sometimes reduced to a few genes. Human chromosome 21 has homologies with at least three mouse chromosomes (10, 16, and 17) and this, as we already mentioned, has hampered the development of mouse models of Down syndrome.

When checked at high resolution, it is sometimes observed that the genes in one species are not exactly in the same order as in another related species, although they are within the same syntenic segment. The genes flanking the *OAS* cluster on human chromosome 12q are on the same syntenic segment as the orthologous genes on mouse Chr 5, but are not in the same order, because a short inversion occurred in one of the lineages (probably in the mouse). Many other such rearrangements have been observed in other regions of the genome (Fig. 5.6).

Based on observations made in several distantly related eukaryotic species, the hypothesis has been suggested (Petkov et al. 2007) that the associations or clustering of genes within short genetic distances might have occurred initially because the genes in question were cooperating in various cellular and physiological functions (akin to large operons, so to speak). It is then not so surprising that these associations have remained relatively unchanged during evolution. Some support for this interesting hypothesis has been provided by the observation of non-allelic parental associations in recombinant inbred strains. Another stronger line of support should come from the analysis of the genome sequence of mice from the Collaborative Cross (see Chaps. 9 and 10).

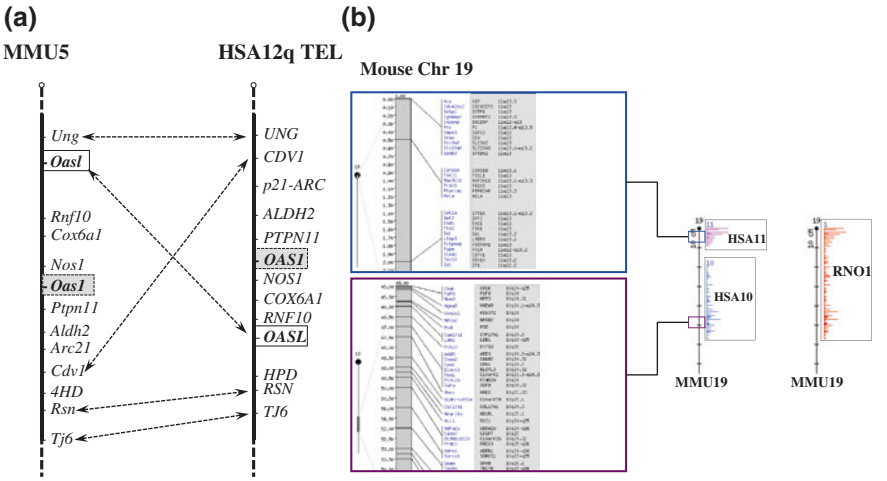

Fig. 5.6 *Homologies of synteny.* **a** An example of homology of synteny between mouse Chr 5 and human Chr 12q24 in the region of the *Oas/OAS* cluster. The genes flanking the *OAS* cluster on human Chr 12q are on the same syntenic segment as the orthologous genes on mouse Chr 5, but not in the same order because a short inversion occurred in the mouse. Many rearrangements of this kind have been observed in other regions of the genome. **b** Another example of homology of synteny between mouse Chr 19 and human Chrs 10 and 11. The same mouse Chr 19 also exhibits homology of synteny with a large fragment of rat Chr 1. More than 90 % of mouse and rat genes are in regions exhibiting homology of synteny with a chromosomal region in humans (the maps are from MGI database—2013)

5.3.5 Gene Families and Pseudogenes

As we already mentioned, when looking in the mouse genome for a DNA sequence orthologous to a human or rat gene we generally find them in the homologous syntenic region, as expected. However, it is not uncommon to find that the sequence homology between the two species is not always in a 1:1 ratio. On human chromosome 12q, for example, there is a cluster of three genes encoding 2′,5'-oligoadenylate synthetase (OAS), an enzyme that is induced by interferons and plays an important role in the inhibition of cellular protein synthesis and resistance towards viral infections. In this cluster, the human genes are arranged in the following order: HSA12 *cen—… –OAS1–OAS3–OAS2– …– tel.*

When looking for the orthologous syntenic region encompassing the OAS encoding genes in the mouse genome, we find a cluster on chromosome 5 with no less than ten genes. These genes exhibit a very high degree of sequence similarity and the linear order: MMU5 *Cen–… –Oas2–Oas3–Oas1e–Oas1c–Oas1b–Oas1f–Oas1h–Oas1g–Oas1a–Oas1d– …– Tel*. Thus, the human *OAS2* and *OAS3* genes each have, and as expected, a single 1:1 orthologous copy on mouse chromosome 5 while the human *OAS1* has no less than eight copies (1:8 orthologs). These *Oas1*

genes are all transcribed although not always in the same direction, indicating that they probably result from a series of segmental duplications with subsequent rearrangements (inversions). In the rat, the structure and organization of the cluster is similar to that of the mouse, but with only eight genes; the orthologous copies of mouse *Oas1a* and *Oas1e* are missing (Perelygin et al. 2006). These differences between the human, rat, and mouse OAS clusters indicate that the genomes of these three species are in constant evolution. Similar observations have been made when performing sequence alignments between mice of the same genus *Mus* but belonging to different species (Fig. 5.7).

These clusters of genes (the three human genes, ten mouse genes and eight rat genes), which encode proteins with similar biochemical functions, were presumably formed by recurrent duplications of a single ancestral gene and represent what geneticists call a *gene family.* Such gene families are common in mammalian genomes and include, for example, the genes encoding the globins, the myosins, the *Hox* and *Sox* clusters, etc. Looking at different unrelated vertebrate species, one observes that the number of repeated copies is highly variable, and the significance of these variations in copy number (if any) is not clear. In the case of the mouse *Oas* cluster, all ten copies are transcribed but the mouse *Oas1b* gene carries a stop codon in its exon 4, resulting in the premature truncation of the

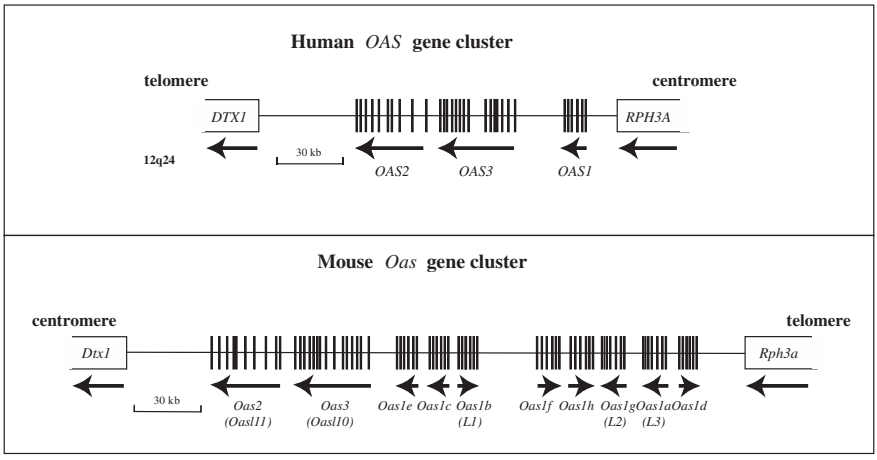

Fig. 5.7 *Gene families.* The three genes encoding human 2′, 5' oligoadenylate synthetase (OAS) are clustered on HSA12, flanked by the same two genes (*DTX1* and *RPH3A*) as in the mouse, and ordered as indicated in the figure. These three genes are transcribed in the same direction. The homologous region is on mouse Chr 5 (MMU5) and consists of ten genes with a very high degree of sequence similarity. The orthologous copies of human *OAS2* and *OAS3* are well preserved, with a 1:1 orthology, while human *OAS1* has no less than eight orthologous copies in the mouse. This cluster of *Oas1* genes probably results from a series of segmental duplication with subsequent rearrangements (inversions). All these genes are transcribed, although not always is the same direction. Such quantitative differences between the human and mouse OAS clusters indicate that the genomes of these species are in constant evolution, although with variations in gene copy numbers (Adapted from Mashimo et al. 2003)

gene product (oligoadenylate synthetase or 2′,5'-OAS), leading itself to its complete inactivation in virtually all mouse laboratory strains. Interestingly, this mutation does not exist in wild mice and researchers demonstrated that this difference, which is specific to the *Oas1b* gene, is responsible for the susceptibility of practically all laboratory mice to experimental infections with flaviruses. The function(s) of the proteins encoded by the other genes of the family is (are) not yet elucidated but, obviously, they do not complement the functional deficiency of *Oas1b* in laboratory strains.

The formation of a gene family results from a mechanism that is classical in evolution. As in the case of the *OAS/Oas* clusters, a majority of these families are formed by a succession of tandem duplications of a single ancestral gene and the different proteins encoded by the genes of the same family (commonly designated *isoforms*) generally have similar biochemical functions, but this is not a rule. Some gene families are easy to identify because the duplicated copies are closely linked to each other, are arranged in tandem, and have retained similar sequences. In other instances the situation is more complex because the gene family is ancient and has been more or less extensively remodeled during evolution. This is the case, for example, with the *Oas1* gene cluster that we described above and two other genes with a similar structure (Oas-like1 and Oas-like2—symbols *Oasl1* and *Oasl2*), located 4 cM away, on the proximal end of the same mouse Chr 5 (Fig. 5.8).

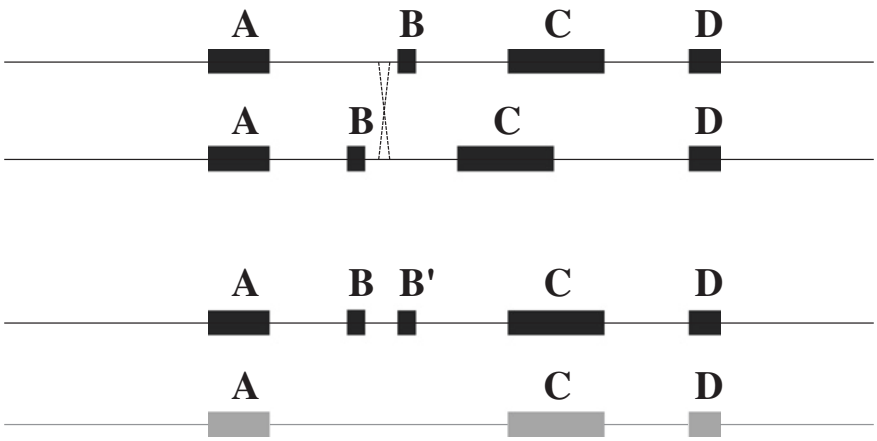

Fig. 5.8 *Gene duplications*. Some tandem duplications result from unequal crossing-over between homologous chromatids, as indicated in the illustration. The chromosome with a deleted gene or segment (in *grey*) is in general rapidly lost, while the duplicated region (*solid black*) is retained. Gene duplications can also result from error (slippage) of the polymerase during DNA replication, with the enzyme copying the same segment more than once. Gene duplication is an essential source of evolutionary innovation when the duplicated copies acquire specific functions (for example, the different isoforms of β-globin, *Hox* genes etc.)

This is also the case for the genes encoding the globin subunits, which are all clearly derived from a single ancestral copy that existed some 500 Myr ago, but are now separated in two different clusters in the mouse genome (α-globin on Chr 11 and β-globin loci on Chr 7). The expansion or contraction of gene families in a specific lineage can be due to chance, or can be the result of natural selection, and it is extremely difficult to decide between these two options.

When genes are duplicated in tandem, it is also common to observe that not all the copies are transcribed in exactly the same way. For example, according to the strain, laboratory mice have either one or two copies of the gene encoding *Renin*, a protein that participates in the regulation of arterial blood pressure (*Ren1* and, sometimes, *Ren2*-Chr 1). *Ren1* encodes the renin mRNAs found in the submaxillary gland while *Ren2* encodes the renin mRNAs found in the kidney. This difference in transcriptional activity can be explained by the promoter regions of these two genes, where structural differences have been described (Panthier et al. 1984).

Some specific gene families, like those concerned with a reproductive function (exhibiting, for example, spermatid or oocyte-specific expression), an immunological function, or an olfactory function (encoding, for example, the odorant (OR) or vomeronasal (VR) receptors) originated from relatively recent duplications (expansions) that occurred in the mouse lineage since the time of its divergence from the rat, around 12–14 Myr ago. In the initial draft of the C57BL/6 genomic sequence, for example, scientists were surprised at the identification of some 1,400 OR genes and 332 VR genes. In the human genome the same olfactory or vomeronasal receptors are much less numerous. The explanation generally proposed to explain these considerable differences is that such sequences are preserved because they are translated into functional proteins that are more or less important for the host species. Geneticists have coined the expression "*genome shaping*" to account for such a situation where the genome structure is influenced by natural selection triggered itself by environmental factors (Nouvel 1994). Although one can accept the idea that olfactory receptors are much more important for wild mice than for human beings, the same argument is less obvious for some other genes that are members of very large gene families in rodents but are much less represented in the human genome.

After careful examination and comparison with a consensus (or ancestral) sequence, it is common to observe that some members of a gene family carry point mutations (SNPs). These mutations are missense or sometimes nonsense, resulting in a loss of function for the gene in question. This is the case for the Oas1-like gene (*Oasl1*) described above. When this occurs, the mutated gene no longer encodes a functional protein, even if it is still transcribed. It is then classified as a *pseudogene* and its sequence will progressively degenerate, generation after generation, until it becomes unrecognizable in terms of structure. The pseudogene is then called a *relic*, a *vestige* or a *fossil*, and the intergenic regions of the genome have sometimes been described as "cemeteries" for these degenerated genes. The "death" of a gene is not important for the survival of the species as long as other copies of the family are present in the genome as potential backups, capable of taking over the function of the missing copies.

When missense mutations (i.e. leading to an amino acid substitution) occur in a gene that is a member of a family, this results in the gene encoding a novel protein, with sometimes new characteristics, a different 3D shape, a different stability, etc. Evolution will then "decide" whether this novel protein deserves to be retained or not based on the potential advantages it may confer to the affected individual in its current environment (Demuth and Hahn 2009). In this case, one realizes that diploid organisms have an advantage since they can put to test, in the same genome and for a few generations, both the ancient and the new copy (allele) of a given gene and finally retain the one with the best fit.

An interesting gene family is that of myosins, mostly known for their role in muscle contraction but also involved in a wide range of motility processes. In fact, myosins belong to a huge superfamily of genes whose products share the basic properties of actin binding, ATP hydrolysis (ATPase enzyme activity), and force transduction. Virtually all eukaryotic cells contain myosin isoforms (alternative forms). Some isoforms have specialized functions in certain cell types (muscle), while others are ubiquitous.

A careful analysis of the initial draft of the mouse genome sequence indicated that, in this species, the rate of nucleotide substitution is approximately twice as fast as the rate in human, and this explains why, after a few million years, it is sometimes difficult to establish sequence similarities between some elements of the human and mouse genomes.

As we discussed, it is clear that the mammalian genome contains a great number of sequences that look like protein-coding genes but, in fact, are not (or no longer). The first category of these sequences is the *pseudogenes* we reported above, which are duplicated copies of an ancestral (single copy) gene, and have become non-functional after the accumulation of random mutations (SNPs or indels). There is, however, another category of pseudogenes that geneticists call *processed pseudogenes*. These pseudogenes, unlike the former ones (which are then called *unprocessed pseudogenes*), originate from the retrotranscription of messenger RNAs back into the genomic DNA in more or less random locations. They have no introns and often exhibit mutations in their sequences (including frame-shifts and stop codons), indicating that they definitely do not encode proteins. Around 18,000 such pseudogenes have been identified in the mouse genome assembly (build 38.1), but their identification is often difficult. To discriminate between a true, *bona fide* gene (a gene encoding a protein and then submitted to purifying selection) and a pseudogene (processed or unprocessed), researchers calculate the so-called K_a/K_s ratio. This ratio compares the number of non-synonymous substitutions (K_a) to the number of synonymous substitutions (K_s) in the sequence of the two genes. Synonymous mutations, as we will discuss later, do not modify the amino acid sequence (for example, the GGC codon becomes GGA but still codes for glycine) and accordingly can occur at the same frequency in genes and in the pseudogenes, with no consequence. Non-synonymous mutations, on the contrary, because they generally alter the protein structure, and often its function, are counterselected and are uncommon in functional genes. Computing the K_a/K_s ratio is then a reliable assessment of whether a gene is a "true gene" or a

pseudogene. K_a/K_s values approaching 1 are indicative of neutral evolution, suggesting a pseudogene. In addition, most mouse pseudogenes do not have an orthologous copy in the same syntenic position in the human or rat genomes, whereas active genes generally do.

As we discussed above, most pseudogenes were considered to be "fossils" or "relics" of genes that, once transcribed and reintegrated into the genome, became silent and functionally useless. This view, however, might not be correct or universally true. In fact, there has been speculation and some evidence has been collected suggesting that pseudogenes, or portions of the latter, may be transcribed from the opposite strand relative to their functional counterparts, making them a source of antisense RNA. These RNAs have been proposed to play a role in the fine regulation of genes of the same family through RNA–RNA interaction (Balakirev and Ayala 2003). Even more recently, scientists working on the mouse transcriptome have identified no less than 10,000 full-length cDNAs derived from expressed pseudogenes—representing approximately 10 % of the known transcriptome—with a good half of them likely participating in various regulatory mechanisms. As noted by the members of the FANTOM 3 project (see later in this chapter), we must remain open-minded about the potential function of expressed pseudogenes. For this reason, pseudogenes have been referred to as "*potogenes*" (potential genes) (Balakirev and Ayala 2003; Hayashizaki and Carninci 2006).

5.3.6 Copy Number Variations

In a preceding section (see 5.3.1.1), while discussing the different structural variations that have been observed at the genome level, we noted that some genes have been found to be missing (deleted) in some mouse strains and not in others (for example, *Snca* on Chr 6), while other genes, in contrast, were duplicated in some strains and not in others (for example, *Ren1* and *Ren2* on Chr 1). Variations of this kind are common in mammals and one can certainly expect many similar cases to be reported in the future, for example when comparing distantly related strains or sub-species of the same *Mus* genus. Many of these duplications are lost after a few generations, but a few of them may be retained, eventually after a few changes, either by chance or because they have an adaptive value. We have already discussed this point.

Copy number variations (CNVs) originate from both coding and non-coding regions of the genome. The mechanisms leading to these CNVs in a specific chromosomal region have not yet been completely elucidated, but it makes sense to consider a priori that CNVs are of three kinds. A substantial proportion probably results from unequal crossovers, producing both deleted and complementary duplicated genomic segments. Given that these chromosomal rearrangements often concern large segments, the duplications have a greater chance to be transmitted to the next generation than the deletions, which are generally unviable and lost.

Another type of CNV probably results from defects occurring during DNA replication (for example, defects in replication fork maintenance). This class of CNV commonly occurs in somatic cell lineages (especially in neoplastic tissues), and, accordingly, occurs independently of the process of meiotic recombination.

Finally, the observation that some short-length chromosomal duplications have been found on different chromosomes (cases have been reported in the mouse) suggests that these duplications are, in fact, transpositions of DNA segments very similar to those described earlier and classified as transcriptionally active pseudogenes.

In the mouse, around 100 well-dispersed regions across the 19 autosomes and the X chromosome have been shown to harbor CNVs. Their greatest preponderance is on chromosomes 7, 12, 14, and X, where some of them appear as large blocks.

The sequence homology between the different copies is >94 % on the average, and their size ranges from 62 bp to 8.6 Mb (with an average length of 250 kb). In total, if we include both the deletions and the duplications, this represents close to 10 % of all polymorphisms (excluding microsatellites), with short deletions being more frequent than insertions (Cutler and Kassner 2008).

CNVs involving large or very large chromosomal segments, although rare, have been observed by cytogeneticists using the classical techniques of fluorescence in situ hybridization. Nowadays, more sensitive techniques, like high-resolution comparative genomic hybridization (HR-CGH) or representational oligonucleotide microarray analysis (ROMA), are adapted to this sort of analysis. Using appropriate DNA arrays, these techniques allow for the detection of structural variations at a resolution of 200 bp (Egan et al. 2007) (Fig. 5.9).

In the near future, taking advantage of the recent advances in DNA sequencing technology, it should be possible to identify and quantify many more CNVs at high resolution in both human and mouse, allowing comparisons to be made at the individual level.

The occurrence of CNVs at the genome level translates to variations in gene dosage within the duplicated or deleted regions (0/1–1/1–2/1, etc.), and it makes sense to think that this may be causative or associated with some pathologies. A trisomic mouse, for example, can be regarded as carrying a single large CNV, since the only difference relative to a normal karyotype is an extra chromosome. This difference can nevertheless result in a severe and often lethal syndrome. A good example where a CNV has been found to be causative of a pathological syndrome is Charcot–Marie–Tooth, type A (CMT1A) disease in humans. This neuropathy was found to segregate with a ~1.4 Mb duplication on human chromosome 17p12 among the members of the same family, suggesting a possible causal relationship. Shortly after this observation, the gene coding for peripheral myelin protein 22 (*PMP22*), a component of myelin, was identified within the duplicated region and mutations in this gene were found to be also responsible for a clinical form of the disease very similar to the form associated with the duplication (Valentijn et al. 1992a, b). Finally, an almost perfect mouse model of CMT1A was created by pronuclear injection of a YAC containing a normal, intact copy of the

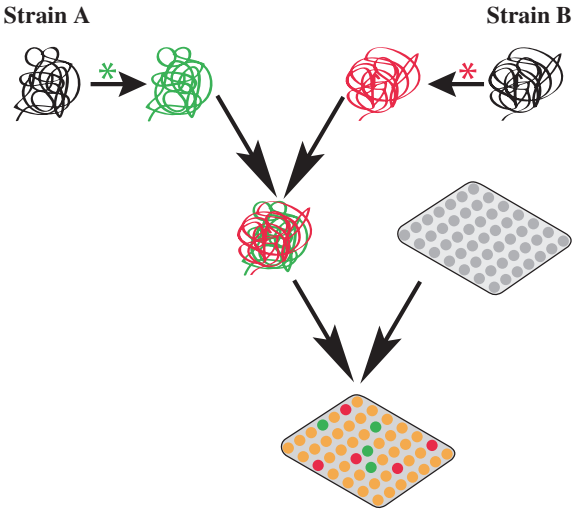

Fig. 5.9 *High-resolution-comparative genomic hybridization (HR-CGH).* This technique is useful for comparing two genomes, or two chromosomal regions, for possible quantitative differences in terms of copy number. The technique consists of two steps. First, a reference DNA sample is labeled with a fluorophore (for example, Cyanine 3, *green*) while the DNA from a test sample is labeled with a different fluorophore (for example, Cyanine 5, *red*). In the second step, equal quantities of the two-labeled DNA samples are mixed and co-hybridized to a DNA microarray of several thousand evenly spaced cloned DNA fragments previously spotted on the array. Finally, after hybridization, digital imaging systems are used to capture and quantify the relative fluorescence intensities of each of the hybridized fluorophores. Obviously, the ratio of the fluorescence intensities is proportional to the ratio of the copy numbers of DNA sequences in the test and reference DNA samples. If the intensities of the fluorophores are equal for a given probe, the spot appears yellow and the region of the genome is interpreted as having an equal quantity of DNA in the test and reference samples (i.e., no copy number variation (CNV)). If there is an altered Cyanine 3:Cyanine 5 ratio, this indicates a loss or a gain of the test DNA sample at that specific genomic region. Discovering which regions of the genome have undergone CNVs is achieved by another test, for example by sequencing followed by fine localization of the sequence. Finely estimating the CNVs can ultimately help to identify genes that are over- or under-expressed, or even deleted. The technique can be adapted to the localization of CNVs directly on the chromosomes

human *PMP22* gene and a large proportion of its flanking region. The conclusions of all these observations and experiments are that both point mutations and duplication of the *PMP22* gene can produce the same phenotype of severe demyelination in the peripheral nervous system.

If the mere duplication of an intact, normal myelin-encoding gene (*PMP22-Pmp22*) can induce a pathology in humans and mice, as demonstrated with YAC transgenics, one can then seriously consider that other CNVs might be at the origin of (or associated with) some clinical diseases or, at least, influence their phenotypic expression (penetrance or expressivity, for example) by altering the transcript level of some essential genes. The presence of some specific CNVs in the human genome has been found to be associated with susceptibility to autism

(Sebat et al. 2007; Cook and Scherer 2008). A reduction in CNVs involving the gene Defensin beta 1 (*DEFB*) has been reported to increase the risk of developing Crohn disease (Roberts et al. 2012). Other human pathologies are equally suspected to be associated with (or the consequence of) CNVs (e.g., autoimmunity, susceptibility or resistance to infectious disease).

In the mouse, genes involved in the control of the immune response or environmental sensory perception have also been found to exist in variable copy numbers in the genomes of the various inbred strains (Watkins-Chow and Pavan 2008). In these conditions, it should not be so surprising to observe in the future that these mice exhibit different phenotypes related to these CNVs.

Nowadays, many geneticists consider that the transmission of some complex traits might be better explained by the transmission of CNVs than by hypothetical Mendelian characteristics (Canales and Walz 2011). Observations relative to some infectious diseases in human populations have already provided preliminary clues to this important question. For example, Gonzalez and colleagues (Gonzalez et al. 2005) reported a strong positive correlation between a high number of copies of the gene encoding the chemokine CCL3L1 and HIV susceptibility.

5.3.7 Single Nucleotide Polymorphisms

When orthologous sequences from different mice (laboratory mice or wild mice) are aligned, single nucleotide differences are frequently observed in the DNA sequence. These differences are base-pair substitutions in most instances, less frequently insertions or deletions of one nucleotide. These sequence differences have been collectively designated *single nucleotide polymorphisms* (SNPs, pronounced "snips") and are the most common type of genetic variation at the DNA level. They are found in both coding and non-coding regions and almost all these SNPs are bi-allelic, i.e., presenting one of two possible nucleotides in an individual (e.g., homozygous G/G or T/T or sometimes heterozygous G/T).

SNPs are extremely abundant among the different mouse inbred strains, and even more so across the different strains recently derived from wild populations. These SNPs are easy to score and permit the performance of high-density/high-resolution mapping. They have undoubtedly been an important outcome of the mouse genome sequencing project, because they represent the ultimate genetic markers. We described their use and advantages in Chap. 4 (Fig. 5.10).

5.3.8 Tandem Repeated Sequences

Like other mammalian genomes, the mouse genome contains a large number of repeated (both coding and noncoding) sequences. They are classified as moderately or highly repeated sequences, and among the latter one must also

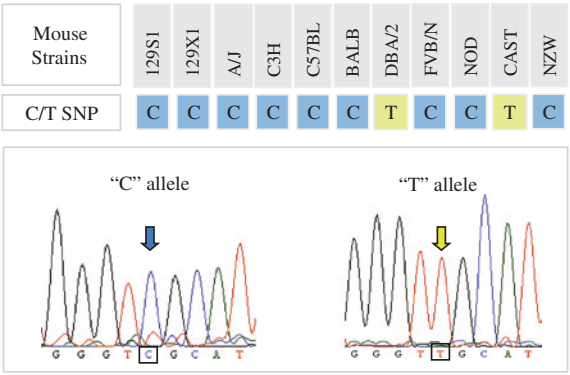

Fig. 5.10 *Single nucleotide polymorphisms (SNPs).* SNPs are single base-pair differences in the DNA sequence, and are the most common type of genetic variation. As described in Chap. 4, they are very useful for genetic mapping, they are found in both coding and non-coding regions, and almost all these SNPs are bi-allelic, i.e., presenting one of two possible nucleotides (e.g., G/G, T/T, or G/T genotypes). In the figure, the upper panel represents a C/T SNP that is polymorphic between DBA/2 and CAST (homozygous for the T allele) and other strains (homozygous for the C allele). The *lower panel* presents DNA sequencing electropherograms showing the SNP (*arrow*)

distinguish those that are organized as tandem repeats and those that are interspersed. *Tandem repeats* are those where the nucleotides motifs are repeated adjacent to each other in a head-to-tail manner. Depending on the number of nucleotides and on the size of the motif, these tandem repeats are known as *satellite* DNA (between 120 and 250 nucleotides), *minisatellites* (between 10 and 60 nucleotides), and *microsatellites* (between 2 and 6 nucleotides). In these types of repeats, the polymorphism is a direct consequence of the number of repeats. The interspersed or dispersed repeats are a totally different category and will be described below.

5.3.8.1 Satellite DNA

The name "satellite DNA" was coined in reference to a difference in the buoyant density of this category of DNA when compared to the density of bulk DNA. Satellite DNA constitutes about 5 % of total mouse DNA and is divided into two major categories: major satellite, which is composed of 234-bp repeats (6 Mb long altogether—occurring at a few loci on the genome), and minor satellite, which is composed of 123-bp repeats (from 500 kb to 1.2 Mb in size and located essentially in the centromeric and telomeric regions of chromosomes). Satellite DNA is the main component of heterochromatin, is not transcribed, and has proved to be rather difficult to sequence.

5.3.8.2 Minisatellites

Minisatellite loci are also known as *variable number of tandem repeats* or VNTRs. They consist of a short series of 10–60 bp repeated in tandem over and over to reach around 5–10 kb in size. They are extremely abundant and are distributed at more than 1,000 locations in mammalian genomes. The occasional slippage occurring during replication is probably at the origin of the minisatellite copy number variations, thereby making each individual unique (Kuznetsova et al. 2005). These highly polymorphic loci were used as genetic markers in the late 1980s, particularly in human studies, and became the basis for the famous DNA fingerprinting method that revolutionized forensic medicine. These "fingerprints" are the individual-specific band patterns resulting from the hybridization (by use of Southern blotting) of restriction-endonuclease-digested DNA with probes directed against extremely polymorphic minisatellite (VNTR) loci. Although it was used in a few mouse linkage studies and also for the genetic monitoring of inbred strains (isogenic individuals within an inbred strain share the same band pattern), the use of DNA fingerprinting in the mouse was abandoned after the advent of microsatellites as universal molecular markers (Julier et al. 1990; Silver 1995).

5.3.8.3 Microsatellites

Microsatellites (also known as short tandem repeats (STRs) or simple sequence length polymorphisms (SSLPs)) are tandem repeats of 1–5-bp elements that are probably the consequence of polymerase slippages. They are very abundant (approximately 10^5 copies per genome), extremely polymorphic, and widely distributed throughout the genome. Since the early 1990s, microsatellites have been the genetic marker of choice in mouse genetics because their analysis is extremely simple, inexpensive, and relatively reliable. For the same reason as for the SNPs mentioned above, we will review their interest as genetic markers in several chapters of this book and in various contexts (Fig. 5.11).

5.3.8.4 Trinucleotide Repeat Expansions

Some severe human genetic disorders have been found to be the consequence of the continuous and abnormal expansion of DNA-trinucleotide repeats in certain genes. The fragile X syndrome is one of these disorders and the first to be explained at the molecular level. Human geneticists found 230–4,000 CGG tandem repeats in a specific X-linked gene in affected patients compared with the 5–54 repeats in unaffected individuals. Similarly, Huntington disease (HD), which affects muscle coordination often associated with psychiatric problems, is caused

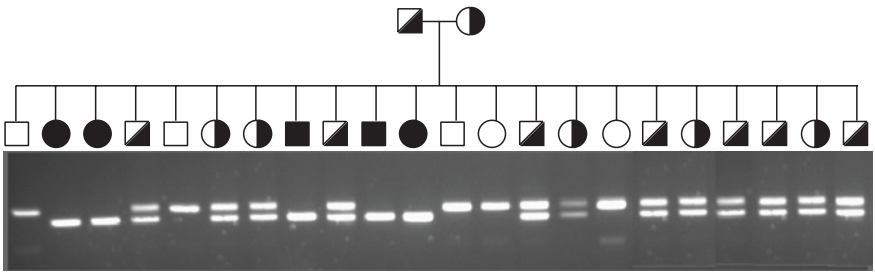

Fig. 5.11 *Microsatellites*. Microsatellites (SSLPs) are composed of short DNA sequences, measuring 1–6 bp, which are repeated in tandem a number of times. They are common in all mammalian genomes, where they exhibit variations in terms of the number of repeats (size polymorphism) and for this reason they are sometimes designated as simple sequence length polymorphisms (SSLPs). Microsatellites can be amplified by PCR with primers designed from the flanking regions. The number of repeats, which translates into size variations of the amplification product, can then be used as a reliable genetic marker. Microsatellites have been extensively used in the mouse for the establishment of high-density/high-resolution genetic maps and are still used for the acute localization of quantitative traits. As indicated in the figure, microsatellites are co-dominant markers, allowing for the identification of heterozygous genotypes. In the figure, we can observe the segregation of the alleles from a microsatellite locus on a pedigree. The male (*square*) and the female (*circle*) from this breeding pair are both heterozygous (*black* and *white*). In the 22 offspring we can clearly see the segregation of the two alleles, where some mice are homozygous (*solid color* on the pedigree) for one allele or the other (one band on the gel), and the rest are heterozygous (two bands). Note that the percentages are in agreement with Mendelian ratios for this co-dominant SSLP marker (~54 % heterozygous; ~23 % homozygous larger allele; ~23 % homozygous smaller allele)

by a CAG repeat expansion in the protein-coding regions of another specific gene called *Huntingtin* (*HTT*—in human Chr 4p16.3). In some other instances, such repeats are also observed and associated with a severe pathology but they are located outside of the protein-coding regions of the genes. To date, similar DNA-trinucleotide repeat expansions have not been reported in the mouse, but transgenic mouse models have been created by pronuclear microinjection of DNAs cloned from affected human patients (Ehrnhoefer et al. 2009).

5.3.9 Interspersed Repeated Sequences: Transposable Elements

Transposable elements (TE), as the name indicates, are small sized DNA sequences that move within the genome and insert into new chromosomal locations sometimes leaving behind a copy of their sequence at their original site (Wessler 2006). These TEs exist in virtually all genomes and have been described in bacteria, *Drosophila*, mammals and many other organisms. TEs were identified

and characterized for the first time in plants, more precisely in maize, through the somatic mutations they induced.[4] In the mouse, and more generally in mammals, these elements are repeated over and over, by thousands of copies, but they are dispersed in the genome and for this reason they are commonly designated *interspersed repeats* in opposition to the *tandem repeats* discussed above. Transposable elements are generally classified into two categories: (i) the *retrotransposons*, which transpose via an RNA intermediate in a "copy and paste" fashion, and (ii) the *transposons*, which use a "cut and paste" mechanism to move within the genome, with no RNA intermediate.

5.3.9.1 The Retrotransposons

Retrotransposons (or class I transposons) are of two kinds based on their size and structure: the LINEs (Long Interspersed Nuclear Elements) and the SINEs (Short Interspersed Nuclear Elements). In addition to these two kinds of transposons, endogenous retroviruses (ERVs) are often considered as equivalent to retrotransposons, as we will explain. Altogether these TEs represent the most abundant component of the mammalian genome, estimated at a proportion of greater than 40 % of genomic DNA.

Long Interspersed Nuclear Elements

The LINE family of retrotransposons, and more precisely the L1 subfamily, is the most important category of transposable elements in placental mammals, representing roughly 17–20 % of mouse genomic DNA. The normal, intact L1 sequence measures ~7.5 kb and consists of a promoter at its 5' end, followed by two non-overlapping open reading frames, ORF1 and ORF2, that encode respectively an RNA-binding protein and a 40-kDa protein with reverse transcriptase and endonuclease activity, and finally an AT-rich region of variable length at its 3' end. This basic structure is relatively uniform, but variations resulting from mutations or deletions, accumulated with time, are common. Thus, only a minority of the LINE elements (a few thousand) appears intact in the mouse genome. The mRNA transcribed from these LINEs serves as templates for the reverse transcriptase II encoded in ORF2, and this explains why this type of transposon is also designated *autonomous transposons*. The new cDNA (a new LINE element) is retrotransposed into a different site, at a new position in the genome, with the help of the endonuclease that nicks the chromosomal DNA and creates the conditions favorable for integration: in other words a true "copy and paste" mechanism. This

[4] Barbara McClintock was awarded the Nobel Prize in 1983 for the discovery of "jumping genes".

process of retrotranscription is similar to the one leading to the creation of processed pseudogenes, as discussed earlier. Sometimes it fails, and this also explains why so many LINEs are incomplete and truncated at their 5' end.

As observed after the sequencing of several mammalian genomes and comparisons between related species, L1 transposons are active as contributors to the so-called genome shaping and have been a source of evolutionary novelty by providing sequence motifs that can be recruited by the host, either for the regulation of its own genes or among its coding sequences. In contrast to this rather positive aspect, L1 transposition can also be deleterious for the host, for example when a transposed copy accidentally inserts within a gene or when it mediates a chromosomal rearrangement through ectopic (non-allelic) recombination (Sookdeo et al. 2013). The *spastic* mutation of the mouse (*Glrb*^{spa}-Chr 3), which is a model of human hereditary hyperekplexia (OMIM 149400), is caused by the intronic insertion of a 7.1-kb L1 element resulting in the aberrant splicing of the beta subunit of the glycine receptor mRNA (Mülhardt et al. 1994). L1 transposition can also be mutagenic in somatic tissues and was actually discovered through this type of activity in maize. This finding has potential consequences for the whole organism which can translate into an increase in cancer occurrences (Belancio et al. 2010). However, most L1 sequences are silenced by methylation and finally become inactive.

This mechanism of LINE retrotransposition, as described, would result in a progressive increase in the size of mammalian genomes unless a compensatory mechanism operates at some point. Based on recent observations, geneticists assume that the mechanism in question consists of repeated deletions (sometimes massive) of some of these constantly burgeoning sequences. Whatever the exact nature of the regulatory mechanism, the size difference observed between the human and mouse genome is generally attributed to variations in the number of L1 copies.

Short Interspersed Nuclear Elements

SINEs are a type of non-autonomous retrotransposon whose sequence does not encode any protein. SINEs have a sequence of around 100–500 bp, which is closely related to the sequence of some tRNAs or to short RNAs. The most common category of SINEs in the human genome is the *Alu*1 sequence, whose equivalents in the mouse genome are the B1 and B2 sequences. SINEs are transcribed by RNA polymerase III but their retrotranscription, necessary for their mobility inside the genome, is not completely elucidated and probably depends (at least in part) upon the LINE machinery—hence their occasional designation as non-autonomous retrotransposons.

There are around 1–1.5 $\times$ 10^6 copies of these SINEs in a mouse genome, representing between 11 and 17 % of the total genomic DNA. Depending on their sequence, SINEs are classified as lineage-specific (added to the mouse genome after the divergence from a common ancestor with other rodents) or ancestral

(before the divergence).[5] Thus, the sequences of these two categories of SINEs have great value for research in evolution and systematics.

Using a software program for multiple sequence alignment guided by phylogenetic trees, researchers have found a DNA sequence measuring 710 bp in the close vicinity of the bovine β-globin locus, sandwiched between two SINEs, and obviously resulting from a transposition (Zelnick et al. 1987). This finding may be considered circumstantial but it nevertheless indicates that, if such a transposition of a DNA segment (by "hitch-hiking", so to speak) can occur in the bovine genome it may also occur in other species, and this is important in the context of the constant remodeling of the genome structure.

The existence of a very large number of retrotransposons with nearly identical sequences, scattered throughout the mouse genome, has some potentially interesting technical applications in the sense that universal (non-specific) primers for PCR amplification can be designed based on the sequence of these retrotransposons and used either with another specific primer (for example, for cloning the sequences flanking a transgenic insertion) or with the same primer with the inverted sequence for the amplification of the host genomic DNA situated between two LINEs or SINEs.

The Endogenous Retroviruses

The *endogenous retroviruses* (ERVs) are a third kind of element that can affect the structure and function of the mouse genome. Although uncommon, infections of mouse germ cells by retroviruses can occur, resulting in the integration of more or less complete retroviral copies into the mouse genome. These retroviral copies are easily recognizable at the molecular level because they are flanked by two classical long terminal repeats (or LTRs) and contain the three classical genes *gag* (encoding structural elements of the virus), *pol* (encoding the reverse transcriptase), and *env* (encoding the coat protein of the virus). Many ERVs are incomplete and no longer move in the mouse genome, and in some cases one LTR is the only sequence that remains of an ancestral retroviral copy that has been completely excised or deleted.

Just like the LINEs and SINEs, ERVs occasionally have influence on the genome's structure and function. They can be mutagenic, like LINEs, when they integrate into the host DNA into or around a coding sequence. They can also trigger various forms of structural rearrangements. A classical example of the role of ERVs as mutagens is the *hairless* mutation of the mouse (Hr^{hr}) (Stoye et al. 1988; Cachon-Gonzalez et al. 1994). This recessive mutation is the result of the retroviral insertion of murine leukemia proviral sequences into intron 6 of a gene encoding a specific protein at the *Hr* locus of chromosome 14, which results in aberrant splicing of the gene. Many other mutations of this type have also been reported in the mouse. The viable yellow (A^{vy}) allele, which originated through the retrotransposition of an

[5] The ancestral SINEs are sometimes designated MIR3 (for mammalian-wide interspersed repeat elements).

intracisternal A-particle[6] (IAP) upstream to the canonical wild-type transcription start of the *agouti* gene (*A*), is another example.

Some elements of these ERVs can also have functional consequences. This is the case, for example, when long terminal repeats (LTRs) act as alternative promoters or enhancers leading to the transcription of tissue-specific RNAs. In humans, diseases have been reported as being caused by TE-generated alleles. These diseases include, for example, hemophilia A and B, severe combined immunodeficiency, porphyria, predisposition to cancer, and some cases of Duchenne muscular dystrophy.

Recombination between homologous retroviral sequences has also contributed to "gene shuffling" and to gene duplications and deletions that largely contribute to genome plasticity.

Several years ago, the retrotransposons we just described were considered as examples of the so-called "selfish" or "junk" DNA because, apparently, their only function was to make more copies of themselves with no apparent benefit for the host. Nowadays, the perspective has dramatically changed and these DNA elements are regarded as tools contributing to genome plasticity and "novelty". L1 sequences frequently insert into the introns of functional genes, where they can interfere with the transcription process without permanently harming the gene product. When the inserted L1 copy is long or very long, the transcription rate is reduced and this might represent another subtle (and reversible) method of gene regulation. When inserted into an intron, SINEs or LINEs can also introduce new splicing sites, allowing the de novo creation of new exons and accordingly of new protein domains. It is then up to the environment to determine, at no risk, whether the new protein presents some selective advantage, whether the structural alteration is selectively neutral or, on the contrary, whether it is detrimental and should be eliminated by returning to the original copy of the gene—which is still in the genome as a back-up. In other words, thanks to the TEs, evolution can perform experiments at virtually no cost.

5.3.9.2 The Transposons

Transposons exist in many species including bacteria, plants, insects (for example the P elements of *Drosophila melanogaster*), and mammals. They are relatively short elements, measuring a few kilobases when intact, and they encode an essential enzyme: a *transposase* (also called *transposonase*). The gene encoding this transposase is flanked by two inverted or palindromic terminal repeats that are essential for transposition in the genome. These terminal repeats pair with each other as the transposon folds and forms a loop. This DNA loop is then excised and released, ready to transpose into another location in the genome, hence the "cut and paste" mechanism of transposition.

[6] IAPs are a class of defective endogenous retroviral sequences measuring ~7 kb. These IAPs are mostly abundant in the endoplasmic reticulum.

The excision of a transposon from its original location in the host genome often generates a small gap in the genomic DNA, while its insertion in a new location disorganizes the neighboring genetic sequences. For these reasons the transposons are responsible for the occurrence of new mutations in the species where they are active.

In the mouse genome the vast majority of transposons no longer encode any functional transposase, and accordingly, they have lost the capacity to transpose: they are "dormant" or even "dead". Interestingly, a fish transposon, which had remained inactive for over 15 million years, could be artificially "resurrected" into an active one by the transgenic addition of two essential functional components into the same host genome: (i) the transposon DNA containing the two inverted terminal repeats, and (ii) the transposase enzyme essential for activation. This engineered (and resurrected) transposon, named *Sleeping Beauty* (Izsvák and Ivics 2005), has been shown to transpose efficiently enough in the mouse to be proposed as a tool for the in vivo production of mutations (Carlson and Largaespada 2005). This method of mutagenesis has the advantage that new mutations are created simply by breeding mice, and, most importantly, that the transposon DNA tags the integration site. However, the disadvantage is that the mutation rate is rather low, especially when compared to other mutagenesis methods. More recently, *SleepingBeauty* has also been reported as an interesting tool for cancer gene discovery and gene therapy (Copeland and Jenkins 2010; Howell 2012), helping for example to introduce transgenes into host genomes. Other resurrected transposons (*Piggy Bac* and *Mariner*) have also been used for the production of mutations (by gene trapping) and for transgenesis.

The transposable elements are definitely important elements of the genome, since they participate actively in its evolution. Together they are often referred to as elements of the "*mobilome,*" and it is likely that their role and functions are still underestimated.

5.4 The Transcriptome: Coding and Non-coding RNAs

In the same issue of the journal *Nature* announcing the initial draft of the mouse genome sequence (*Nature* 420–5 December 2002), another very important report was published, summarizing the results of the functional and manual annotation of a large collection (60,770) of full-length mouse cDNA[7] collected by the "FANTOM consortium" (Functional Annotation of the Mouse) of the RIKEN Genomic Science Center in (Okazaki et al. 2002). This publication, perhaps because it was released at the same time as the impressive and outstanding

[7] Full-length cDNA libraries are established from all RNA transcripts (protein-coding and non-protein-coding). Manual annotation of such libraries is a guarantee of their quality.

presentation of the mouse genome sequence, did not receive the attention we think it deserved from the community, at least when published. Ten years later, and based on the information gathered in the meantime from the analysis of the mouse and human genomes and transcriptomes, we think that this report should be considered another breakthrough in our understanding of the ways in which the mammalian genome actually works. Not only did it confirm some important observations that were made independently a few years earlier, for example about the unjustified overestimation of the number of protein-coding genes in the mouse genome (which was sometimes estimated to be as high as 120,000) and the concomitant underestimation (or mis-appreciation) of some other transcription products (Lander et al. 2001; Kapranov et al. 2002), but it also raised a number of new ideas that have been confirmed since and widely amplified in successive reports, in particular those of the same FANTOM consortium as well as in other reviews devoted to the analysis of the mouse transcriptome (Carninci et al. 2005; Katayama et al. 2005; Mattick and Makunin 2006; Gustincich et al. 2006; Saxena and Carninci 2011; ENCODE Project Consortium 2012; Kapranov and St Laurent 2012). The ideas that were developed in these initial reports have radically changed our views of the transcriptome, in particular the belief which was solidly anchored in most scientists' mind that proteins were the most important (if not the only) bioactive molecules encoded in the genome.

The main conclusions of the reports in question are the following: (i) the protein-coding RNAs (the mRNAs) and the other RNAs that cooperate with mRNAs in protein synthesis and processing (rRNAs, tRNAs, snoRNAs, and snRNAs) represent only a minor (around ~2–3 %) component of the transcriptome; (ii) the mouse genome is pervasively and extensively transcribed and encodes several thousand non-protein-coding RNAs (ncRNAs), and (iii) sequencing all these RNA molecules and making in silico alignments with the DNA genomic sequence indicates that up to 90 % of the euchromatic genome of the mouse is transcribed, sometimes from both DNA strands, and in both directions (many sense–antisense pairs).

Nowadays, the mammalian genome can no longer be regarded as a mere repository of the basic information necessary for the synthesis of thousands of proteins, but rather as a sophisticated factory releasing a great variety of coding and non-coding RNAs (ncRNAs) of various sizes and functions. In spite of enormous progress in the sequencing technology of nucleic acids, the inventory of these molecules is far from being completed and their annotation may still require several years. It has been established, for example, that many primary RNA transcripts are processed into smaller sized molecules, while others are alternatively spliced, thus tremendously increasing the complexity and diversity of the transcriptome. For this reason, scientists sometimes refer to this new category of non-coding RNAs as "*the dark matter of the transcriptome*". We will summarize the situation as it stands presently based on recent reviews on the subject, but it is clear that this chapter, more than any others in this book, will require regular updating. Undertaking the exhaustive inventory of the ncRNAs encoded in the mouse genome and performing their annotation is nothing less than embarking on the exploration of "*a new continent in the RNA world*".

5.4.1 ncRNAs Involved in Protein Synthesis

In addition to the messenger RNAs (mRNAs), which are protein-coding and are considered as the "noble" RNAs since they represent the message transcribed from the DNA, four types of ncRNAs have been described as essential components in the successive steps of protein synthesis and processing: transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), short non-coding RNAs (snRNAs, sometimes referred to as U-RNAs) and small nucleolar RNAs (snoRNAs).

5.4.1.1 Transfer RNAs

Transfer RNAs are a relatively homogeneous family of "adaptor" molecules whose function is to mediate recognition of a specific codon in the processed mRNA and to provide the corresponding amino acid during the process of protein synthesis (elongation step). These RNAs are part of a larger transcript, the pre-tRNAs, in the nucleus, which are subsequently split into smaller molecules (average 80 nucleotides), with a typical 3D cloverleaf structure that is well adapted to the function since, as we said, the tRNA binds to the mRNA codon (specific), to the new incoming amino acid (specific), and to the last amino acid incorporated into the growing polypeptidic chain. There are around 500 tRNA-encoding genes in the mouse genome and about the same number of pseudogenes. The tRNA-encoding genes are dispersed over the whole genetic map, including both the X and the Y chromosomes. Their computerized prediction is difficult due to their short size and, mostly, to the existence in the mouse genome of a very high number of short interspersed sequences (SINEs—see above) that originally derived from tRNA genes but are now inactive. This is typically a source of confusion and, for this reason, the experimental verification of the status of a potentially novel tRNA gene must be part of the annotation process (Coughlin et al. 2009).

5.4.1.2 Ribosomal RNAs

In contrast with the tRNAs, ribosomal RNAs are a relatively heterogeneous family of molecules with a size between 150 and ~5,000 nucleotides. The family comprises four types of RNAs (28S, 5.8S, 5S, and 18S). The 28S RNA (5,070 nt) and the 5.8S RNA (156 nt) bind to each other and are associated with the 5S RNA (121 nt) and with at least 45 proteins, to make the ribosomal large unit (60S). The 18S rRNA (comprising 1,869 nt) is associated with around 33 proteins to make the ribosomal small unit (40S). The two ribosomal subunits, the small and the large, are tightly associated to make the cytoplasmic ribosomes. The biosynthesis of mature ribosomes is complex and involves numerous processing events with the participation of other ncRNAs. When mature, the ribosomes serve as workbenches for protein synthesis. The mRNA is held sandwiched between the two subunits of

the rRNAs while being "scanned" and then transcoded into proteins. rRNAs are rapidly degraded in the cytoplasm once they have been used for protein synthesis. The genes encoding ribosomal RNAs are very numerous and spread over the whole genome (Henderson et al. 1974). They are organized in repeated units that, in the mouse, are 44 kb long. Each repeat contains three of the genes encoding rRNA, namely 18S, 5.8S, and 28S, and constitutes a transcription unit producing polycistronic RNA that is cleaved apart afterwards. These units are tandemly repeated and constitute the so-called nucleolar organizers (or NORs). These are distributed over several chromosomes (Chrs 4, 12, 15, 16, 18 and 19) in the case of *Mus m. domesticus*, but on all 40 chromosomes except the Y in *Mus caroli* (Rowe et al. 1996; Cazaux et al. 2011). At the end of mitosis (telophase) when rDNA transcription by RNA Polymerase I resumes, the NORs gather in the nucleolus (a nuclear organelle where rRNAs are produced and assembled with ribosomal proteins to form functional ribosomes). Genes that encode rRNA are expressed in virtually all types of cells and in all species, including prokaryotes. For this reason, many rRNAs have been sequenced and their sequences are now used as tools for systematics (ribotyping).

5.4.1.3 Small Nuclear RNAs

Small nuclear RNA molecules are found in the nucleus of eukaryotic cells. As is the case for many other small-sized RNAs, they are transcribed as larger molecules that are cleaved afterwards. They have an average length of approximately 150 nucleotides and are generally classified into five categories: U1, U2, U4, U5, and U6. Each of these snRNAs is associated with a large set of specific proteins (over 150), and the complexes they form with these proteins are referred to as small nuclear ribonucleoproteins (snRNPs or "snurps"). The snurps are essential in the splicing process. The splicing of mRNAs is a very complex and extremely precise process and this is probably why the spliceosome requires so many components to make its functioning totally error-proof. Each of the five categories of snRNAs has specific binding sequences and a specific function on the pre-mRNA substrate.

5.4.1.4 Small Nucleolar RNAs

The small nucleolar RNAs are small molecules measuring 60–300 nt. They are involved in the processing of rRNAs and are essential for ribosome maturation. They can also regulate the splicing of some mRNAs by modifying small nuclear RNAs (snRNAs) that are the major RNA component of the spliceosome, as we mentioned. snoRNAs probably have many other functions that have not yet been described, and the inventory of this family of molecules is difficult because their computerized prediction and classification is unreliable, yielding many orphan snoRNAs. snoRNAs encoding genes have been identified at several loci in the

mouse genome (2, 7, 8, 9, 12, 17, and X). The range of functions of these RNAs is likely to expand with the discovery of new molecules (Gardner et al. 2010).

Some genetic diseases affecting humans (for example spinal muscular atrophy and congenital dyskeratosis) have been correlated to abnormal functioning of the snurps. Prader–Willi syndrome (and the reciprocal Angelman syndrome—see Chap. 6 for details) is caused by the abnormal imprinting of a cluster of snoRNAs encoding genes located in the q11-13 region of human chromosome 15 that are involved in the synthesis of the serotonin-2C receptor mRNA. snRNAs also play an important role in maintaining the size of the telomeres (see Chap. 3).

5.4.2 The ncRNAs Functioning as Post-transcriptional Regulators

5.4.2.1 MicroRNAs

MicroRNAs (miRNAs) are small, single-stranded RNAs, measuring 21–24 nt (average 22 nt), whose function is to negatively regulate specific genes by mRNA degradation or translational repression. Around 60 % of these miRNAs are encoded in the intergenic regions and in antisense orientation to certain genes, and 40 % are encoded in the intronic regions of genes encoding proteins. These RNAs (along with the small interfering RNAs, described later) are the most well-known family of non-protein-coding RNAs.

The DNA encoding miRNAs is transcribed into precursors called pri-miRNAs. Each of these pri-miRNAs folds to form a double-stranded structure by base-pairing with itself. This structure looks like a hairpin with a few loops of stranded RNA. The pri-miRNA is then cleaved into a precursor known as a pre-miRNA, which is transported into the cytoplasm. Finally, the pre-miRNA is incorporated into a molecular complex of proteins of the argonaute family called the *miRNA-induced silencing complex* or *miRISC*. The processing of mature miRNAs requires the participation of an endoribonuclease known as *Dicer* that cleaves the pre-miRNA into the mature miRNA. miRISC modulates the activity of the targeted mRNA by identifying a 2–7-bp complementary sequence, known as the "*seed region*", which is generally located at the 3'-UTR. Both the processing and the loading of miRNAs into the RISC complex and the function of this machinery are precisely regulated (Ebert and Sharp 2012).

The fact that these miRNAs exist in several species including invertebrates and plants, and the way they are transcribed and processed from highly preserved sequences, with highly sophisticated mechanisms, indicates that they probably represent an ancestral mechanism of gene regulation (Lewis and Steel 2010). Because they also have a wide range of spatial and temporal expression patterns, they probably play important roles at different steps of embryonic development and in some pathological conditions. Indeed, it is expected that about 60 % of mammalian protein-coding genes are more or less regulated by miRNAs.

miRNAs are numerous and distributed throughout the genomes of both animals and plants. In the mouse, as in humans, their number has been estimated in the range of 1,000. miRNAs are involved in many regulation processes, including cell proliferation, differentiation, apoptosis, and development. They function via base-pairing with complementary sequences of mRNA molecules (seed region), leading either to translational repression or to silencing via target degradation.

miRNA nomenclature consists of the generic or root symbol *Mir*, followed by the numbering in the miRBase database (www.mirbase.org), a database that tracks microRNAs reported for all species. Mouse *Mir143* (microRNA 143), for example, is represented as mmu-mir-143 in miRBase, with the mmu signifying *Mus musculus* (Fig. 5.12).

Demonstration of the involvement of miRNAs in a given developmental or pathological process is not easy. In the mouse, this can be achieved, for example, by performing the complete elimination of all miRNAs in a certain tissue or cell type and then observing the phenotypic effects. Since the *Dicer* protein is essential for the processing of miRNAs, as discussed above, mice with a conditional knockout allele of *Dicer* targeted in Purkinje cells (see Chap. 8—targeted knockout) no longer had any miRNAs in these cells, and were found to develop ataxia with Purkinje cell degeneration. This indicates that at least some miRNAs are indispensable for the differentiation of these highly specific cells (Schaefer et al. 2007). Another more specific strategy would be to establish an indisputable causal and direct relationship between a point mutation in the sequence of a given miRNA and a particular phenotype. Examples of this type are now accumulating,

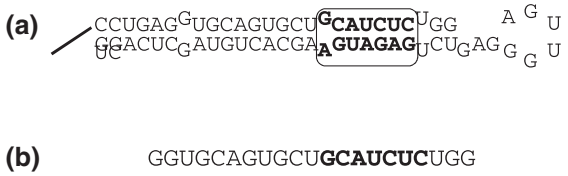

Fig. 5.12 *The microRNAs.* MicroRNAs (miRNAs) are short, noncoding, single-stranded RNAs. These miRNAs are nested within longer non-coding RNA molecules, which are processed in several successive steps with a double-stranded pre-miRNA (**a**), and finally a functional single-stranded RNA molecule measuring 20–22 bp (**b**). These miRNAs finely regulate the expression levels of several genes by binding to the 3'-untranslated regions of the corresponding mRNAs. The seed sequence of miR-143 (represented in *bold*) matches perfectly with the 3'-UTR of the mRNA transcribed from the (cytosine-5)-methyltransferase 3A (*DNMT3A*) gene. mir-143 is known to be involved in cardiac morphogenesis, it has also been implicated in human colon cancer development, and its expression is down-regulated during mouse odontoblast differentiation. mir–143 is encoded in mouse Chr 18 and is transcribed from the same DNA as another miRNA (mir-145). miRNAs are highly conserved in vertebrates, and this is suggestive of an important function. It is expected that about 60 % of mammalian protein-coding genes are more or less regulated by miRNAs

and one of the first and most well-documented cases is the semidominant mutation *Diminuendo* (symbol *Mir96*^{Dmdo}-Chr 6) (Lewis et al. 2009; Lewis and Steel 2010). This mutation was observed in the progeny of a male treated with the chemical mutagen ENU (see Chap. 7) and was presumably induced by this substance. The phenotype is characterized by progressive deafness, a condition that is quite common in humans. After positional cloning and careful sequencing of several candidate DNA segments in the 4.96-Mb critical interval where *Diminuendo* was mapped, the researchers finally found an A →T transversion in the "seed" region of the miRNA *Mir96*. This mutation, which was unique to *Diminuendo* and absent in all other mice as well as in a large series of vertebrates, was confirmed as the causative agent of the deafness and was associated with the down-regulation of several (at least five) proteins, each of them being involved in the function of the hair cells of the inner ear. These five proteins, which are downstream in the cascade of regulation initiated by *Mir96*, are all important for the differentiation and function of the hair cells and were all found to result in deafness when individually knocked out.

The discovery of the molecular origin of the *Diminuendo* mutation is an example of the role that the myriad of miRNAs may play in the fine regulation of gene (mRNA) expression in several developmental or pathological processes in vertebrates. The discovery of a point mutation in the seed region of *Mir96* proved that cell differentiation and organogenesis involve a network of functionally linked proteins as well as one or several miRNA(s).

Identification of the miRNA targets would certainly represent an enormous step forward in developmental genetics, and this is therefore a focus in many laboratories worldwide. Progress, however, is hampered by the fact that miRNAs are very small molecules and their sequences are not often totally complementary to their targets. In addition to this difficulty, many scientists also believe that many mRNAs, if not all, offer several targets to several miRNAs in their 3'-UTRs, thus adding even more complexity to the picture.

MicroRNAs definitely have a promising future in medicine because they are simple molecules but have, at the same time, the power of interfering with gene regulation. In humans they are intensively studied because their expression levels have been found increased in certain forms of cancers (for example, lymphomas or chronic lymphocytic leukemias), in diseases like cardiomyopathies, and in some infectious diseases or autoimmune diseases. These increases in specific miRNAs can then be used as information for the diagnosis or prognosis of the disease, or as potential treatments. For example, aortic banding in mice induces cardiac hypertrophy and concomitant up-regulation of many (over 100) miRNAs including *Mir21*. When *Mir21* was knocked down using an antisense approach, cardiomyocyte hypertrophy was reduced, suggesting that this particular miRNA plays a key role in the mechanism of cardiac hypertrophy. This obviously opens perspectives for the development of novel therapies.

Scientists believe that there are different grades in the process of mRNA regulation by miRNAs. Some miRNAs regulate specific individual targets, but it

seems that key miRNAs (so-called "super-miRNAs") can regulate the expression levels of hundreds of genes simultaneously and cooperatively. These super-miRNAs are of course actively searched. It has been suggested that miRNAs exert both absolute and fine-tuned control of gene expression, adjusting levels of transcripts to give either complete repression or simply decreased expression. Such "fine-tuning" miRNAs will be much harder to identify than those resulting in the complete "switching off" of a gene, since loss of function of any of these miRNAs would presumably have subtle effects, which would be difficult to characterize and study.

The discovery over the past ten years of these post-transcriptional regulators has opened up a "*new continent of the RNA world*". We just gave a rapid overview of this continent using the miRNAs as examples, but many other RNA or RNA-like molecules are just as interesting. We will now consider the case of siRNAs, another type of ncRNA with post-transcriptional regulatory functions.

5.4.2.2 Small Interfering RNA

Small interfering RNA, short interfering RNA, or silencing RNAs (all abbreviated siRNAs) are short double-stranded RNA molecules (20–25 bp) with a 2-bp 3' overhang and phosphate groups on the 5' end of each strand. These RNAs interfere with (i.e., reduce or suppress) the expression of specific genes with complementary nucleotide sequence, and in so doing they obviously have similarities in their mode of action with the miRNAs discussed above.

The existence of these siRNAs and their remarkable properties were discovered by chance while plant geneticists were performing transgenic experiments with the aim of darkening the color of petunia flowers. The transgene they were using was that for chalcone synthase, a key enzyme of the flavonoid/isoflavonoid biosynthesis pathway. The scientists expected that by increasing the enzyme level with several extra transgenic copies of the gene, this may influence the pigmentation of the flower (Napoli et al. 1990). In fact, and to their surprise, instead of obtaining the dark purple flowers they expected they got light-colored flowers and sometimes flowers with white (unpigmented) patches, indicating that the chalcone-encoding transgene actually had adverse effects on the pigmentation process. Other similar experiments revealed that the observed phenotypes were not exceptional but, on the contrary, the consequence of an increased rate of mRNA degradation leading to specific gene suppression or, more precisely, down-regulation. This effect was designated RNA interference or RNAi.

In 1998, Fire and colleagues (1998), performing a similar experiment with the worm *Caenorhabditis elegans,* concluded that neither the complete mRNA nor a variety of antisense RNAs had an effect on protein production in experimentally injected worms. However, they found that double-stranded RNAs corresponding to a myofilament protein successfully silenced the targeted genes,

once injected under the same conditions. They also demonstrated that only a few molecules of injected double-stranded RNA were required to induce gene silencing, thus arguing against stoichiometric interference with endogenous mRNA and suggesting that there could be a catalytic or amplification component in the interference process. This finding had a great impact in biology and medicine when it was demonstrated that RNAi mechanisms are universal and active in humans as well as in several model organisms including rats and mice, offering new tools for gene annotation as well as opening the way to the development of novel therapeutic strategies for the treatment of genetic diseases, including cancers.[8]

Unlike in many model species, RNA interference cannot be triggered in mammalian cells by injecting long double-stranded RNAs, because the cells recognize these RNAs as viruses and immediately develop a deleterious interferon response with consequences for cell survival. Short molecules do not trigger this reaction when injected into the cells.

siRNAs can also be synthesized as single-stranded molecules in the laboratory and then introduced into cells either by direct injection or by transfection. Direct chemical synthesis has the great advantage of allowing slight variations in the sequence, and as a result increasing the efficiency of the siRNAs. Not all native siRNAs are equally active, and the possibility of synthesizing novel molecules appears to be a promising strategy (Ramachandran and Ignacimuthu 2013). The mechanisms by which miRNAs and siRNAs work are similar. However, while miRNAs cause translational repression or destabilization, the siRNAs cleave their target RNAs at a particular site.

The use of RNA interference is an interesting and efficient way of altering the gene function and accordingly of performing gene annotation. However, in most instances and unlike other strategies described in Chap. 8, RNA interference induces down-regulation of gene expression (knockdown) and not knockout proper. In addition, some of these knockdowns are not specific.

5.4.2.3 Piwi-Interacting RNAs

Piwi-interacting RNAs (*piRNAs*) are short ncRNAs (26–31 nt long), which are expressed mainly, not to say specifically, in spermatogenic cells of mammals. Their function is not yet fully understood, but it is known that they form complexes with the regulatory piwi (or miwi) proteins. These piRNA complexes are thought to play a role in transposon silencing in male germ line cells, limiting the expansion of these repeated sequences. They presumably have other functions that have not yet been characterized.

[8] A. Fire and C. Mello were awarded the Nobel Prize in Physiology or Medicine in 2006 for their discovery of "RNA interference—gene silencing by double-stranded RNA".

5.4.2.4 Long Non-coding RNAs

Long non-coding RNAs (lncRNAs) have an average size larger than 200 nt and in many cases, in the range of 2 kb or more. This relatively great size distinguishes them from all other ncRNAs, but being similar in size to the mRNAs can hamper their isolation and characterization. Computer algorithms assessing the coding potential of the two molecules (lncRNAs and mRNAs) have been used to discriminate between these molecules when necessary, but this criterion has finally proven unreliable because some (not all) lncRNAs do have a coding frame or, more precisely, a nucleotide sequence resembling a coding frame with start and termination codons. So far, the analysis of the sequences of lncRNAs does not allow sorting them in discrete families with specific functions. In addition, the sequences of these RNAs are only poorly conserved across species, even among closely related mammals. Indeed, this family of ncRNAs is heterogeneous to the point where its very existence has long been debated. Since lncRNAs are four times more numerous than mRNAs, one can understand why they have been designated the "dark matter" of the transcriptome.

Aside from this rather confusing situation, some data have recently emerged that make the situation a little more coherent. First, sequence alignments reveal that lncRNAs are transcribed from both strands and in both directions overlapping introns, sometimes exons, and intergenic regions: this is never the case with mRNAs. Also, unlike mRNAs, many of these molecules stay in the nucleus, suggesting that they have a function at or close to this location. Finally, and as we will discuss further, the density of lncRNAs seems to be locally associated with some pathologies, suggesting that they may be involved more or less directly in these processes.

Most of the knowledge we have of the lncRNAs results from the studies of five important lncRNAs that have been studied in the mouse and whose functions have now been relatively well characterized: these are the *Kcnq1* overlapping transcript 1 (*Kcnq1ot1*-Chr 7), the antisense IGF2R-RNA (*Airn*-Chr 17), the HOX transcript antisense RNA (*Hotair*-Chr 15), the X-specific transcripts (*Xist*-Chr X), and the X (inactive)-specific transcript, antisense (*Tsix*-Chr X). The function and mode of action of the lncRNAs involved in the X-chromosome inactivation process will be analyzed in Chap. 11. *Xist* is one of the first genes, expressed after fertilization, leading to silencing of all the genes on the targeted chromosome as a consequence of histone H3 modifications. The targeting of XIST RNA to only one of the chromosomes is controlled by another lncRNA: TSIX, which is the antisense repressor of *Xist* on the active X chromosome.

Antisense repression is also the mode of action of the gene *Kcnq1,* whose expression is silenced by the paternally expressed antisense non-coding RNA KCNQ1OT1.

LncRNAs have extremely variable stability and expression levels. Some have a half-life of only one hour (for example, KCNQ1OT1), while others are much more stable. Some are highly expressed, while others are barely detectable.

Indeed, from the many reviews that have been published, one can conclude that "*we have barely begun to scratch the surface of the lncRNA world*" (Kung et al. 2013).

5.5 Ultraconserved Elements (UCE) and Long Conserved Non-coding Sequences

When the mouse genomic sequence is aligned to the genomic sequence of other vertebrate species, we observe that quite a large number of elements measuring ≥200 bp are conserved, and sometimes highly conserved. These sequence elements are commonly designated *ultraconserved elements* (UCEs). UCEs were first described in the human, rat, and mouse genomes by Bejerano and coworkers (2004), but were also discovered in many other more distantly related species (chicken, for example). For the UCEs encoding proteins or functional RNAs, geneticists have an explanation: they consider that these resemblances are a consequence of strong selection pressures acting during evolution and that we mentioned earlier as "genome-shaping forces". However, the situation is much less clear for the non-protein-coding UCEs, and in this case explanations are lacking.

After alignment of the mouse and human genomes, scientists at the RIKEN Institute identified over 600 such conserved non-coding DNA sequences with nearly 95 % identity and a size greater than 500 bp, most of them independent of the previously reported UCEs (Sakuraba et al. 2008). These sequences, which they provisionally designated *long conserved non-coding sequences* (LCNS), were also found scattered throughout the genome of the rat as well as other vertebrate species (chick, frog, fish) but were not found in non-vertebrate species. Given that the probability of finding sequence similarities of that kind, just by chance, is extremely low, two hypotheses were proposed by the researchers to account for their observations: the first hypothesis was to consider that these LCNS either have an important although unknown function associated with their structure (they could have regulatory or structural elements important for the chromosome structural organization, for example), or that they are transcribed into functional ncRNAs whose function is not yet established (perhaps a type of lncRNA); in both cases, this would explain why the sequences in question were selectively constrained. The second hypothesis is that the LCNS/UCEs have remained intact for so many years of evolution, simply because they are mutational cold spots (Katzman et al. 2007). To challenge these hypotheses, the scientists had the clever idea of performing ENU mutagenesis and measuring, afterwards, the frequencies of induced mutations in the LCNS and comparing it with other genomic regions. They did not find any significant differences in the mutation rates after screening 40.7 Mb of conserved sequences (~35 mutations) and concluded that the LCNS were not mutational cold spots. To date, we do not have any satisfactory explanation to account for the presence of so many of these LCNS/UCEs. The scientists of the ENCODE project consider them to be associated with gene regulation (ENCODE Project 2012) and their role is probably essential if we consider their near-universal conservation across extremely divergent species. On the other hand, it has also been reported that deletions of these UCEs in mice had virtually no effect on the viability or fertility of the animals (Ahituv et al. 2007). This indicates

that extreme sequence constraints do not necessarily correspond to crucial functions. For mouse geneticists, this also indicates that another type of sequence element must now be added to the "*dark matter of the transcriptome*".

5.6 Mitochondrial DNA

Mitochondria have a genome of their own that is represented by a small, circular, double-stranded DNA molecule known as mtDNA, sometimes mDNA. In the mouse (as in humans) there are between two and ten such mtDNA molecules per mitochondrion, and the number of mitochondria per cell is extremely variable and depends on the cell type. The oocyte, for example, contains up to 10^6 mtDNA copies while a mature sperm cell contains less than 100.

The mtDNA comprises 37 genes encoding 13 mitochondrial enzymes involved in respiration and oxidative phosphorylation, two ribosomal RNAs (12S and 16S) and a full set of 22 tRNAs that are essential for the synthesis of these enzymes. However, this small set of proteins represents only a sampling of the ~1,500 mitochondrial proteins, the rest of them being encoded in the nuclear genome. In contrast to the mammalian nuclear DNA, mtDNA is a naked DNA molecule (i.e., histone-free) with no introns and no sequence repeats (Bayona-Bafaluy et al. 2003). In addition, its two strands are quite different from those in the nuclear DNA, the heavy strand being very heavy while the light one is much lighter. All these unique characteristics of the mtDNA molecule are generally correlated with its presumptive evolutionary origin, which states that the mitochondria are remnants of bacteria that have been incorporated into the primitive eukaryotic cells and retained as symbiotic organisms due to their selective advantages for cellular metabolism. This interesting hypothesis, which is also proposed for chloroplasts in plants, is not formally confirmed but it seems more than likely and fits perfectly with the molecular data accumulated recently, in particular some fundamental changes in codon usage[9] (Yu et al. 2009).

The consensus sequence of the mouse mtDNA has been established and found to consist of around 16,300 bp, with point variations (a few SNPs and gaps or indels) among the most common laboratory inbred strains and the most commonly used mouse species (Goios et al. 2007, 2008). These sequence polymorphisms have been cleverly exploited to establish or to confirm the phylogenetic relationships relationships between the different species of the genus *Mus* and related genus (see Chap. 1) and the historical phylogeny among the laboratory strains (see Chap. 9). This has allowed, in particular, the confirmation that a great majority

[9] There are a few differences between the vertebrate mtDNA code and the "universal" code. In the mtDNA, UGA codes for Trp rather than being a stop codon. In the same mtDNA there are two Met codons (AUA and AUG) rather than only one. Finally, both AGA and AGG are read as stop codons.

of the most frequently used inbred strains were all derived from the same female ancestor, as initially established by Yonekawa et al. (1982), and to confirm that most laboratory strains can be sorted into three groups with independent ancestral/geographical origins: the Sino-Japanese mice, the Swiss mice and the "Abbie Lathrop's" mice in the United States.

The mtDNA replicates at a much higher rate than the nuclear DNA and does not possess repair mechanisms as efficient as those of the latter. For this reason, and probably also because the mtDNA is not protected from the mutagenic action of its environment by a variety of histone proteins, as is the case for mammalian DNA, it is more "mutable" and appears to be about 10–20 times more affected by mutations generating a sequence polymorphism than the nuclear DNA of the same species. Considering the great differences between male and female gametes in terms of mitochondria numbers (up to 1/1,000), it is no surprise to learn that the mtDNA is transmitted by the mother to her offspring rather than by the father. Although sperm cells do have some mtDNA molecules, the mtDNA appears to be lost very early during egg development, and in virtually all species studied so far the only mtDNA molecules found in embryos are of maternal origin.

When a mutation occurs in a mtDNA molecule of an oocyte (or of a precursor cell), it is generally counter-selected and rapidly eliminated unless it confers a selective advantage to the mtDNA, for example by increasing its replication rate (Sharpley et al. 2012). In the latter case, the mutant molecules progressively overgrow the population of normal mtDNAs and the oocyte (or cell) becomes heteroplasmic with two (or more) types of mtDNA. Finally, due to some sort of sampling effect, sometimes referred to as a genetic bottleneck, the mutant form of the mtDNA may completely replace the pre-existing form and become the standard. This explains why mtDNA is an attractive molecule to geneticists studying evolution. It is also interesting to note that mtDNA evolution is completely independent of nuclear DNA evolution, and accordingly represents another valuable tool for establishing the systematics of a species. For this reason, it has been extensively used in many domestic species, including the mouse, and still is.

In the human species, mutations in the mtDNA have been associated with more or less severe pathologies. Leber hereditary optic neuropathy (LHON), for example, was the first reported and is one of the most prevalent, with an estimated frequency of 15 in 100,000 births. This syndrome is the consequence of mutations (several have been described) occurring in the genes encoding the oxidative phosphorylation complex I. Many other mtDNA defects have been reported in the human species, including a syndrome of maternally inherited diabetes and deafness (MIDD), Leigh syndrome, a syndrome associating neuropathy, ataxia, retinitis pigmentosa, and ptosis (NARP), myoneurogenic gastrointestinal encephalopathy (MNGIE), and many other neuromuscular diseases. All these pathologies are maternally transmitted and exhibit variations in severity presumably associated with the degree of heteroplasmy. Surprisingly, no such pathologies clearly attributable to an mtDNA defect have ever been reported in the mouse, although mtDNA mutations have been reported in cell lines transplanted in vitro.

Because spontaneous mtDNA defects have never been reported in the mouse, animal models of human pathologies have been created by introducing defective human mtDNAs into mouse oocytes.[10] In particular, a murine model of LHON syndrome has been produced by using this strategy. These mice exhibited reduction in retinal function, indicating that the physiopathology of the syndrome may result from some oxidative stress (Lin et al. 2012).

Those readers of this chapter who might be interested in the biology and pathology of mtDNA, in both human and mouse, should refer to the important contribution of D.C. Wallace (University of Pennsylvania), who wrote several reviews on the subject (Wallace 2009).

5.7 Conclusions

At the beginning of this chapter we stated that we considered the decision taken several years ago to completely and systematically sequence the mouse genome to be a wise one. If we consider the huge amount of information gathered, directly or indirectly, from this sequencing and the data we can expect to collect in the near future, our initial feeling is strengthened; indeed, the sequencing of the mouse genome has had an enormous impact in many areas of genetics and biology.

The knowledge of this sequence has allowed the development of better tools (for example, SNPs) and allows better experiments to be designed. Nowadays one can design an experiment of homologous recombination (targeted mutagenesis) with precision at the base-pair level.

Aside from these technical advances, in silico comparisons of the mouse sequence with other mammalian (or vertebrate) sequences has allowed the discovery of similarities or differences that have proved a rich source of information for a better understanding of evolution. Even within individuals of the same species, the analysis of copy number variations, for example, has revealed intriguing differences whose significance and phenotypic expression is not yet completely clear, even if we suspect that they probably play an important role in quantitative genetics.

The information gathered concerning the structure of the mouse genome and its variations across the different inbred strains and different subspecies of the *Mus* genus will certainly reveal important clues for understanding the genetic determinism of complex traits, especially when complemented by the constantly increasing amounts of phenotyping data. The mouse is unique in the sense that one can cross animals of different subspecies, breed very large progenies, extensively phenotype all the animals, and sequence the individual genomes when desired.

[10] Two inbred strains of mice with the same genomic (nuclear) DNA but different mtDNAs are said to be conplasmic. The production of such strains can be achieved by normal sexual reproduction or by direct cytoplasmic transfer (See Chap. 9).

The sequencing of the genome has also revealed its great plasticity. We now know that LINEs and ERVs play an important role in gene regulation, and even as a source of diversity, a point that was totally unexpected.

Finally, a true revolution in our understanding of the transcriptome occurred during the last ten years. The number of protein-encoding genes has been revised downward while the number of RNA-encoding genes is constantly being revised upward. Over the last ten years we have started to realize that a myriad of ncRNAs (long and short) are transcribed from the genome, exhibiting great although still incompletely explored functional diversity. From whole-genome analyses using microarrays and high-throughput transcript sequencing, we estimate that more than 85 % of the nucleotides in the euchromatic genome are represented in primary transcripts. Indeed, the proportion of supposedly "junk" DNA shrinks more every day. We have learnt that the genome is pervasively and bidirectionally transcribed, increasing tremendously the amount of information that can be stored. The discovery of the role of miRNAs and siRNAs in the fine regulation of gene activity is another revolution that may have major consequences for the diagnostic and treatment of some diseases. The long coding RNAs probably play a major role in gene regulation and imprinting … but we have information about only a handful of these molecules although we know that there are many.

The role and importance of the ultraconserved elements and long conserved non-coding sequences remains a mystery. If they are ultraconserved this would mean that they are under selection pressure. But, alternatively, we know that they can be experimentally deleted with apparently no consequences. No doubt all these observations will fuel much research in the years to come and it won't be surprising that, at this point, even the concept of gene may be reconsidered[11].

Acknowledgements The authors thank Doctor Benoît Robert, Institut Pasteur, for his contribution to Sect. 5.3.3 of this chapter.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM (2007) Deletion of ultraconserved elements yields viable mice. PLoS Biol 5:e234

Arnold CN, Xia Y, Lin P, Ross C, Schwander M, Smart NG, Müller U, Beutler B (2011) Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis. Genetics 187:633–641

Balakirev ES, Ayala FJ (2003) Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet 37:123–151

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the splicing code. Nature 465:53–59

[11] Most of the data provided in this chapter concerning the Mouse Genome are from the Ensembl website: http://www.ensembl.org/Mus_musculus/Info/Annotation#assembly and http://www.ncbi.nlm.nih.gov/projects/mapview/stats/BuildStats.cgi?taxid=10090&build=38&ver=1.

Bayona-Bafaluy MP, Acín-Pérez R, Mullikin JC, Park JS, Moreno-Loshuertos R, Hu P, Pérez-Martos A, Fernández-Silva P, Bai Y, Enríquez JA (2003) Revisiting the mouse mitochondrial DNA sequence. Nucleic Acids Res 31:5349–5355

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325

Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P (2010) Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res 38:3909–3922

Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, Clarke SL, Wenger AM, Nguyen L, Gurrieri F, Everman DB, Schwartz CE, Birk OS, Bejerano G, Lomvardas S, Ahituv N (2012) Coding exons function as tissue-specific enhancers of nearby genes. Genome Res 22:1059–1068

Blanco E, Guigo R (2005) Predictive methods using DNA sequences—analysis at the nucleotide level. In: Baxevanis AD, Francis Ouellette BF (eds) Bioinformatics: a practical guide to the analysis of genes and proteins, 3rd edn. Wiley, Hoboken

Buckler AJ, Chang DD, Graw SL, Brook JD, Haber DA, Sharp PA, Housman DE (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc Natl Acad Sci U S A. 88:4005–4009

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Cachon-Gonzalez MB, Fenner S, Coffin JM, Moran C, Best S, Stoye JP (1994) Structure and expression of the hairless gene of mice. Proc Natl Acad Sci U S A 91:7717–7721

Canales CP, Walz K (2011) Copy number variation and susceptibility to complex traits. EMBO Mol Med. 3:1–4

Carlson CM, Largaespada DA (2005) Insertional mutagenesis in mice: new perspectives and tools. Nat Rev Genet 6:568–580

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626–635

Cazaux B, Catalan J, Veyrunes F, Douzery EJ, Britton-Davidian J (2011) Are ribosomal DNA clusters rearrangement hotspots?: a case study in the genus Mus (Rodentia, Muridae). BMC Evol Biol 11:124. doi:10.1186/1471-2148-11-124

Cook EH, Scherer SW (2008) Copy-number variations associated with neuropsychiatric conditions. Nature 455:919–923

Copeland NG, Jenkins NA (2010) Harnessing transposons for cancer genes discovery. Nat Rev Cancer 10:696–706

Coughlin DJ, Babak T, Nihranz C, Hughes TR, Engelke DR (2009) Prediction and verification of mouse tRNA gene families. RNA Biol 6:195–202

Cutler G, Kassner PD (2008) Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. Cytogenet Genome Res 123:297–306

Demuth JP, Hahn MW (2009) The life and death of gene families. BioEssays 31:29–39

Ebert MS, Sharp PA (2012) Roles for microRNAs in conferring robustness to biological processes. Cell 149:515–524

Egan CM, Sridhar S, Wigler M, Hall I (2007) Recurrent DNA copy number variation in the laboratory mouse. Nat Genet 39:1384–1389

Ehrnhoefer DE, Butland SL, Pouladi MA, Hayden MR (2009) Mouse models of Huntington disease: variations on a theme. Dis Model Mech 2:123–129

ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 4:e1001046

ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Fire A, Xu S, Montgomery M, Kostas S, Driver S, Mello C (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806–811

Gardner P, Bateman A, Poole AM (2010) SnoPatrol: how many snoRNA genes are there? J Biol 9:1–4

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521

Goios A, Pereira L, Bogue M, Macaulay V, Amorim A (2007) mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res 17:293–298

Goios A, Gusmão L, Rocha AM, Fonseca A, Pereira L, Bogue M, Amorim A (2008) Identification of mouse inbred strains through mitochondrial DNA single-nucleotide extension. Electrophoresis 29:4795–4802

Goldberg ML. 1979. PhD Diss. Stanford University, Stanford, CA

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440

Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, Lazarevic D, Hayashizaki Y, Carninci P (2006) The complexity of the mammalian transcriptome. J Physiol 575:321–332

Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet 13:469–483

Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigó R (2009) Identifying protein-coding genes in genomic sequences. Genome Biol 10:201 Epub

Hatzis P, van der Flier LG, van Driel MA, Guryev V, Nielsen F, Denissov S, Nijman IJ, Koster J, Santo EE, Welboren W, Versteeg R, Cuppen E, van de Wetering M, Clevers H, Stunnenberg HG (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Mol Cell Biol 28:2732–2744

Hayashizaki Y, Carninci P (2006) Genome Network and FANTOM3: Assessing the Complexity of the Transcriptome. PLoS Genet 2(4):e63

Henderson AS, Eicher EM, Yu MT, Atwood KC (1974) The chromosomal location of ribosomal DNA in the mouse. Chromosoma 49:155–160

Hill RE (2007) How to make a zone of polarizing activity: insights into limb development via the abnormality preaxial polydactyly. Dev Growth Differ 49:439–448

Hoskins AA, Moore MJ (2012) The spliceosome: a flexible, reversible macromolecular machine. Trends Biochem Sci 37:179–188

Howell VM (2012) Sleeping beauty—a mouse model for all cancers? Cancer Lett 317:1–8

International Human Genome Sequencing Consortium, Lander E, Linton L, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) Initial sequencing and analysis of the human genome. Nature 409:890–921

Izsvák Z, Ivics Z (2005) Sleeping Beauty hits them all: transposon-mediated saturation mutagenesis in the mouse germline. Nat Methods 2:735–736

Julier C, de Gouyon B, Georges M, Guénet JL, Nakamura Y, Avner P, Lathrop GM (1990) Minisatellite linkage maps in the mouse by cross-hybridization with human probes containing tandem repeats. Proc Natl Acad Sci U S A. 87:4585–4589

Kapranov P, St Laurent G (2012) Dark matter RNA: existence, function, and controversy. Front Genet. 3:60

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–919

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium (2005) Antisense transcription in the mammalian transcriptome. Science 309:1564–1566

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D (2007) Human genome ultraconserved elements are ultraselected. Science 317:915

Kozak M (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol 196:947–950

Kung JT, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. Genetics 193:651–669

Kuznetsova IS, Prusov AN, Enukashvily NI, Podgornaya OI (2005) New types of mouse centromeric satellite DNAs. Chromosome Res 13:9–25

Lagha M, Bothma JP, Levine M (2012) Mechanisms of transcriptional precision in animal development. Trends Genet 28:409–416

Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhattar R (2013) Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature 494:497–501

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A. 99:7548–7553

Lewis MA, Steel KP (2010) microRNAS in mouse development and diseases. Semin Cell Develop Biol 21:774–780

Lewis MA, Quint E, Glazier AM, Fuchs H, De Angelis MH, Langford C, van Dongen S, Abreu-Goodger C, Piipari M, Redshaw N, Dalmay T, Moreno-Pelayo MA, Enright AJ, Steel KP (2009) An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. Nat Genet 41:614–618

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei CL, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, Fullwood MJ, Cheung E, Liu E, Sung WK, Snyder M, Ruan Y (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148:84–98

Lin CS, Sharpley MS, Fan W, Waymire KG, Sadun AA, Carelli V, Ross-Cisneros FN, Baciu P, Sung E, McManus MJ, Pan BX, Gil DW, Macgregor GR, Wallace DC (2012) Mouse mtDNA mutant model of Leber hereditary optic neuropathy. Proc Natl Acad Sci U S A. 109:20065–20070

Mashimo T, Glaser P, Lucas M, Simon-Chazottes D, Ceccaldi PE, Montagutelli X, Desprès P, Guénet JL (2003) Structural and functional genomics and evolutionary relationships in the cluster of genes encoding murine 2′,5'-oligoadenylate synthetases. Genomics 82:537–552

Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15 Spec No 1:R17–29

Modrek B, Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34:177–180

Mouse ENCODE Consortium (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13:418

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Mülhardt C, Fischer M, Gass P, Simon-Chazottes D, Guénet JL, Kuhse J, Betz H, Becker CM (1994) The spastic mouse: aberrant splicing of glycine receptor beta subunit mRNA caused by intronic insertion of L1 element. Neuron 13:1003–1015

Munroe R, Schimenti J (2009) Mutagenesis of mouse embryonic stem cells with ethylmethanesulfonate. Methods Mol Biol 530:131–138

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J et al (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 296:1661–1671

Napoli C, Lemieux C, Jorgensen R (1990) Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. Plant Cell 2:279–289

Nouvel P (1994) The mammalian genome shaping activity of reverse transcriptase. Genetica 93:191–201

Ohno S (1972) So much "junk" DNA in our genome. In: Smith HH (ed) Evolution of genetic systems. Gordon and Breach, New York, pp 366–370

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I et al (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573

Panthier JJ, Dreyfus M, Roux TL, Rougeon F (1984) Mouse kidney and submaxillary gland renin genes differ in their 5' putative regulatory sequences. Proc Natl Acad Sci U S A. 81:5489–5493

Perelygin AA, Zharkikh AA, Scherbik SV, Brinton MA (2006) The mammalian 2′-5' oligoadenylate synthetase gene family: evidence for concerted evolution of paralogous Oas1 genes in Rodentia and Artiodactyla. J Mol Evol 63:562–576

Perez CJ, Dumas A, Vallières L, Guénet JL, Benavides F (2013) Several classical mouse inbred strains, including DBA/2, NOD/Lt, FVB/N, and SJL/J, carry a putative loss-of-function allele of Gpr84. J Hered 104:565–571

Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K (2007) Evidence of a large-scale functional organization of Mammalian chromosomes. PLoS Biol 5(5):e127

Ramachandran PV, Ignacimuthu S (2013) RNA interference-a silent but an efficient therapeutic tool. Appl Biochem Biotechnol 169:1774–1789

Roberts RL, Diaz-Gallo LM, Barclay ML, Gómez-García M, Cardeña C, Merriman TR, Gearry RB, Martin J (2012) Independent replication of an association of CNVR7113.6 with Crohn's disease in Caucasians. Inflamm Bowel Dis 18:305–311

Rowe LB, Janaswami PM, Barter ME, Birkenmeier EH (1996) Genetic mapping of 18S ribosomal RNA-related loci to mouse chromosomes 5, 6, 9, 12, 17, 18, 19, and X. Mamm Genome 12:886–889

Sakuraba Y, Kimura T, Masuya H, Noguchi H, Sezutsu H, Takahasi KR, Toyoda A, Fukumura R, Murata T, Sakaki Y, Yamamura M, Wakana S, Noda T, Shiroishi T, Gondo Y (2008) Identification and characterization of new long conserved noncoding sequences in vertebrates. Mamm Genome 19:703–712

Savarese F, Grosschedl R (2006) Blurring cis and trans in gene regulation. Cell 126:248–250

Saxena A, Carninci P (2011) Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. BioEssays 33:830–839

Schaefer A, O'Carroll D, Tan CL, Hillman D, Sugimori M, Llinas R et al (2007) Cerebellar neurodegeneration in the absence of microRNAs. J Exp Med 204:1553–1558

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung

W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M (2007) Strong association of de novo copy number mutations with autism. Science 316:445–449

Sharpley MS, Marciniak C, Eckel-Mahan K, McManus M, Crimi M, Waymire K, Lin CS, Masubuchi S, Friend N, Koike M, Chalkia D, MacGregor G, Sassone-Corsi P, Wallace DC (2012) Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. Cell 151:333–343

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B (2012) A map of the cis-regulatory sequences in the mouse genome. Nature 488:116–120

Silver LM (1995) Mouse genetics—concepts and applications. Oxford University Press, Oxford

Sookdeo A, Hepp CM, McClure MA, Boissinot S (2013) Revisiting the evolution of mouse LINE-1 in the genomic era. Mob DNA 4:3. doi:10.1186/1759-8753-4-3

Specht CG, Schoepfer R (2001) Deletion of the alpha-synuclein locus in a subpopulation of C57BL/6 J inbred mice. BMC Neurosci 2:11

Stoye JP, Fenner S, Greenoak GE, Moran C, Coffin JM (1988) Role of endogenous retroviruses as mutagens: the hairless mutation of mice. Cell 54:383–391

Valentijn LJ, Baas F, Wolterman RA, Hoogendijk JE, van den Bosch NH, Zorn I, Gabreëls-Festen AW, de Visser M, Bolhuis PA (1992a) Identical point mutations of PMP-22 in Trembler-J mouse and Charcot-Marie-Tooth disease type 1A. Nat Genet 4:288–291

Valentijn LJ, Bolhuis PA, Zorn I, Hoogendijk JE, van den Bosch N, Hensels GW, Stanton VP Jr, Housman DE, Fischbeck KH, Ross DA et al (1992b) The peripheral myelin gene PMP-22/GAS-3 is duplicated in Charcot-Marie-Tooth disease type 1A. Nat Genet 1:166–170

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291:1304–1351

Wallace DC (2009) The pathophysiology of mitochondrial disease as modeled in the mouse. Genes Dev 23:1714–1736

Watkins-Chow DE, Pavan WJ (2008) Genomic copy number and expression variation within the C57BL/6 J inbred mouse strain. Genome Res 18:60–66

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci U S A. 103:17600–17601

Wong K, Bumpstead S, Van Der Weyden L, Reinholdt LG, Wilming LG, Adams DJ, Keane TM (2012) Sequencing and characterization of the FVB/NJ mouse genome. Genome Biol 13:R72

Xia Y, Won S, Du X, Lin P, Ross C, La Vine D, Wiltshire S, Leiva G, Vidal SM, Whittle B, Goodnow CC, Koziol J, Moresco EM, Beutler B (2010) Bulk segregation mapping of mutations in closely related strains of mice. Genetics 186:1139–1146

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Migita S, Bonhomme F, Hjorth JP, Petras ML, Tagashira Y (1982) Origins of laboratory mice deduced from restriction patterns of mitochondrial DNA. Differentiation 22:222–226

Yu X, Wester-Rosenlöf L, Gimsa U, Holzhueter SA, Marques A, Jonas L, Hagenow K, Kunz M, Nizze H, Tiedge M, Holmdahl R, Ibrahim SM (2009) The mtDNA nt7778 G/T polymorphism affects autoimmune diseases and reproductive performance in the mouse. Hum Mol Genet 18:4689–4698

Zelnick CR, Burks DJ, Duncan CH (1987) A composite transposon 3' to the cow fetal globin gene binds a sequence specific factor. Nucleic Acids Res 15:10437–10453