1
Deciphering DNA Sequence Information

Mark Kaganovich and Michael Snyder

1.1
Introduction

The revolution in DNA sequencing technologies during the past decade and a half has resulted in an outburst of genome sequence information for more than 800 organisms. Genomes of many humans from different ethnic backgrounds have been sequenced at varying degrees of coverage using multiple technological platforms and strategies and the effort is ongoing; thousands of human genomes will be available in the next few years for researchers to analyze. A major challenge ahead is to determine the functional components of the different genome sequences and how they vary across individuals and species.

Traditionally most efforts have focused on the analysis of protein-coding genes. These are typically annotated as exons separated by introns. Genes are transcribed into messenger RNA (mRNA), the introns are spliced out, and the exons are translated to protein. In addition we now know there is a plethora of information in non-coding DNA sequence as to how to regulate the expression of the gene-coding regions. In this chapter we cover the major categories of genomic sequence and the methods used to investigate them.

1.2
Genes and Transcribed Regions

Genes are transcribed regions of the genome that are made up of exons and introns. Exons are arranged linearly on the transcribed portion of the DNA separated by introns. The entire region is transcribed, the introns are spliced out by the cellular splicing machinery, a poly-A tail is added to the 3′ end of the RNA, and a modified guanine is added to the 5′ end, termed the 5′ cap. The resulting mRNA is exported from the nucleus into the cytoplasm and translated.

Genome Organization And Function In The Cell Nucleus, First Edition. Edited by Karsten Rippe.

Published 2012 by WILEY-VCH Verlag GmbH & Co. KGaA

1.2.1 Open Reading Frames

It is thought that the human genome is made up of roughly 20 000 distinct protein-coding genes, though this number is greatly increased when considering the many protein-coding combinations that result from alternate splicing of introns (Chapter 11). This means that if exons A, B, and C make up a gene, two isoforms could be the exons A and B spliced together and the exons A and C spliced together. The average human exon length in humans is 140 bp. The median intron length is $\sim$1000 bp [1]. The average is approximately 3000 bp, due to the long tail of the intron lengths distribution [1]. There are some introns that are greater than 100 000 bp and $<$10% are longer than 11 000 bp [1, 2]. Recent evidence suggests that >90% of human genes have alternate splicing isoforms that are spliced in a tissue-dependent manner [3, 4]. The exons in a gene that are spliced together into an mRNA and then translated to protein are referred to as an open reading frame (ORF), distinguished by the often species-specific start and stop codons.

Mapping transcribed regions of the genome and the ORFs contained within them is an important challenge in genomics. It is necessary for our fundamental understanding of cellular function; we cannot understand the cell without knowing its protein components that are coded for by genes. Recent work in our laboratory on high-throughput sequencing of the model system yeast *Saccharomyces cerevisiae* transcriptome has helped reveal the complex nature of ORF organization in eukaryotic cells [4, 5]. Nagalakshmi *et al*. sequenced and quantified the transcriptome of *S*. *cerevisiae* (under rich media conditions) by capturing and reverse-transcribing poly-adenylated mRNA and then fragmenting and sequencing the cDNA (using the Illumina high-throughput sequencing platform). The resulting 35-bp sequencing reads were mapped to the genomic sequence. The mapping is thought to represent much of the transcribed region of the genome, not including non-polyadenylated RNAs such as microRNAs (miRNA) and ribosomal RNA (rRNA). Since polyadenylation is a requirement for mRNA export from the nucleus (with the exception of histone mRNAs), these regions should include all translated ORFs. Some of the RNAs are likely non-coding regulatory RNAs that are also polyadenylated, such as long interfering non-coding RNAs (lincs) [6].

1.2.2 Mapping Transcriptional Start Sites

Sequencing transcribed RNA (RNA-Seq) has helped map the location of transcriptional start sites in the genome, which is integral for our understanding of transcriptional promoter structure and thus gene expression regulation [3, 5]. The yeast genome includes many overlapping transcripts transcribed from opposite strands of the DNA [5]. Because many of these are antisense it is expected that they form double-stranded RNA (dsRNA) species. This phenomenon is likely also present in mammalian genomes, which is surprising given the role of dsRNA in

triggering the RNA interference machinery in many eukaryotes that silences gene expression and in potentially triggering viral immune responses [7, 8].

Furthermore, recent evidence suggests that the majority of yeast transcriptional promoters are inherently bidirectional [9, 10]. Some of the opposite strand transcripts of ORF-containing mRNAs are immediately degraded by the cell whereas others are stable, though most do not code for protein [9, 10]. Their function is not fully understood, though regulatory roles have been suggested [9, 10] that are discussed in Chapter 13). Thus, much of the yeast genome is transcribed (74.5% of the non-repetitive sequence), but only a fraction of the genomic sequence is translated to protein [5]. The same seems to be true for mammalian genomes [2, 7, 8]. Some genomic regions code for "non-coding" or not translated RNA such as: miRNA, small nucleolar RNA (snoRNA, involved in translation), transfer RNA (tRNA, involved in translation), rRNA (compose the ribosome), piwi-interacting RNA (piRNA, retrotransposon silencing), and long interfering non-coding RNA (lincs, regulatory; Chapters 13 and 14). The recent discoveries of regulatory non-coding RNAs has significantly augmented the earlier understood biological paradigm that dictated that the function of RNA was to act as an intermediate between the DNA code and protein assembly.

Further mapping of genomic transcriptional start sites has been attempted via chromatin immunoprecipitation (described in more detail later in this chapter) with anti-RNA Polymerase II (Pol II) antibodies. Pol II is phosphorylated on its C-terminal domain when it is activated for transcriptional initiation. Thus, Pol II binding in its hypophosphorylated state is a likely indicator of a transcriptional start site [11]. Previous work correlating Pol II human genomic binding positions with microarray-measured gene expression suggested that up to a third of genes that are not expressed are bound by Pol II in the promoter region [12]. However, more recent RNA-Seq analysis of the human transcriptome has identified transcripts from these promoters, thus potentially repudiating the earlier model in which Pol II associated with silent promoters [3].

1.2.3
Mapping Untranslated Regions on mRNA

Untranslated regions (UTRs) are regions of coding mRNAs transcripts that are not translated. UTRs are present both on the 5′ and 3′ ends and are significant for regulation of mRNA translation, localization, and degradation. In higher eukaryotes 3′ UTRs are targets of miRNAs that facilitate mRNA translation silencing either through RNA degradation, inhibition of translation initiation, or mRNA sequestration. Transcriptome sequencing has helped to map UTRs by allowing a comparison of the transcribed RNA sequence and the predicted ORF (predicted by species-specific start and end codons) [5, 8].

Approximately half of all human and mouse transcripts have alternative start codons 5′ of the main ORF start site, defined as an alternate, new out of frame ORF, termed an upstream ORF (uORF) [13]. Recent work has suggested that genes with uORFs are expressed at lower levels, possibly as a result of competing

ribosome binding sites on the mRNA [13]. The level of mRNA transcription is not affected by uORF presence but translation is significantly lower. Thus, the arrangement of ORFs on the mRNA, in addition to 5′ and 3′ UTRs, represent another layer of gene expression regulation. Polymorphisms that either disrupt or introduce a uORF into a gene have been linked to the occurrence of various human diseases such as melanoma and hereditary pancreatitis [13].

Much is still unknown about the exact location of the translational start codons in relation to the transcriptional start sites on the DNA. The presence of complex overlapping transcripts and the difficulties associated with genome-wide mapping of short sequencing reads has presented challenges to effectively annotating distinct transcriptional start and end sites and the ORFs associated with the transcripts. Improvements in sequencing and mapping technology will likely help resolve these genomic ambiguities. Longer read lengths will make transcript sequence identification more accurate and single molecule sequencing RNA will help avoid biases introduced by retro transcription and cDNA amplification. Recently, the Helicos single molecule sequencing technology has been used to sequence RNA molecules directly [14]. This approach promises to better sequence and quantify a wider range of transcripts in terms of both length and abundance.

1.3 Non-Coding Genomic Elements

1.3.1 Pseudogenes

Genome annotation is complicated by the presence of pseudogenes, which are genetic sequences that resemble genes but do not code for functional protein. Pseudogenes are most often thought of as evolutionary remnants of functional genes whose function was "lost" either because of redundancy with other genes, such as would occur if they were born out of gene duplication, or simply as a result of changing cellular needs. Pseudogenes created through gene duplication are referred to as non-processed pseudogenes. Processed pseudogenes are those that are created through retrotransposition in which an RNA transcribed from a gene is reverse transcribed and inserted into the genome [15]. Further, both mechanisms have the potential to also generate new functional genes.

Pseudogenes present a challenge to proper genome annotation because of their sequence similarity to functional genes. Computational methods have been attempted to properly distinguish them [16]. Pseudogenes in the genomic regions characterized as part of the *enc*yclopedia *o*f *D*NA *e*lements (ENCODE) have approximately 60% sequence identity to their parent gene [16]. They seem to have the same mutation rates as neutral DNA sequence, which is to say they are not subject to the evolutionary constraints of that of protein coding genes or other functional elements. Nevertheless, there is abundant evidence that at least a portion of pseudogenes are transcribed, though not translated [16]. This may

suggest a functional role for at least a subset of pseudogenes. One indirect function of some pseudogenes is their role in non-allelic homologous recombination (where paralogous sequences recombine during meiosis) [17]. This phenomenon is responsible for several disease phenotypes in humans, due to the resulting gene deletion or duplication [17]. For instance, Gaucher disease results from recombination between the gene for β-glucosidase and a neighboring pseudogene [17].

1.3.2 Repeats

As much as 45% of the human genome is thought to be composed of sequence repeats of various lengths and frequencies [18, 19]. The shortest in size are simple repeats or short tandem repeats, also referred to as microsatellites, which are 1–10 bp. They are likely the result of DNA polymerase slippage errors during replication in which the polymerase-nascent strand complex splits back relative to the template, thus repeating previously covered template sequence [20]. Repeats of length 10–60 bp are called minisatellites, and longer ones are tandem repeats. Tandem repeats are found at the centromere and telomere positions of chromosomes where they are thought to carry out a structural role [19]. Together, this class of repeats represents 1.81% of the genome [19].

Segmental duplications (SD) refer to large-scale repeats that result from duplications of large segments of the genome and have been preserved throughout genome sequence evolution. SDs cover approximately 12–15% of the human genome and can be up to 630 kb in size; this is based largely on studies mapping human variation in the presence or absence of large genomic segments, referred to as copy number variants (CNVs) [21–23]. Earlier estimates of 15% seem to be inflated because of lower resolution mapping of CNVs, thereby increasing the genomic segments classified as duplications [21, 24]. SDs are defined as genomic segments that map to multiple locations on the genome with $\geq 90\%$ sequence identity. Given the assumption of neutral evolution as a measure of the rate of divergence of genomic sequence, 90% sequence identity corresponds to 35–40 million years of evolution since the duplication event creating the SD [24]. Interestingly, the human genome has a far greater number of SDs than other mammalian genomes of comparable length. The mouse genome is 6.6% SD and the chimpanzee genome is thought to be $\sim 4.8\%$ [24]. This points to the unique complexity of the human genome. Furthermore, SDs in humans and chimpanzees differ from those in the mouse and rat genomes in that 48% of human SDs are thought to be interchromosomal (the duplications are located on different chromosomes) whereas in mice it is 13% and in rats 15% [24].

There are three main categories of SDs: pericentromeric SDs are found near the centromere, subtelomeric SDs are found near the telomere regions of chromosomes, and interstitial SDs which are found in between the telomere and centromere regions. Duplicates in subtelomeric regions seem to have their origin in the subtelomeric region of the same chromosome or, more commonly, in other chromosomes. Pericentromeric and interstitial SDs tend to originate from either region.

Interestingly, SD distribution among chromosomes is not random for pericentromeric and interstitial SDs; some chromosomes are more likely than others to contain SDs. One of the most highly correlated properties of the density of duplicated regions on a chromosome is gene density. In other words, the occurrence of lots of genes in a region is a predictor of SD presence, albeit a weak one [24]. This suggests an evolutionary role for SD in the formation of new genes through duplication.

SDs result in new gene formation, disruption of genes, and CNVs between species and within species. As a consequence, SDs, and the reciprocal outcomes of the recombination events that form SDs (deletions), are correlated with a large number of genetic diseases [17]. Disease causing CNVs often involve highly dosage-specific genes.

1.3.3 Structural Variants

Large-scale genomic rearrangements including SDs and all other possible permutations of >1 kb DNA stretches that vary between individual genomes are referred to as structural variants (SVs). SVs include duplications, inversions, deletions, insertions, and translocations. CNVs are a particular example of SVs, and the most widely studied variation so far [21, 22, 24, 25]. Previously, it was thought that much of the individual variation among humans is due to single nucleotide polymorphisms (SNPs), and that individuals were only about 0.1–1.0% divergent by sequence [21]. However, more recent work suggests that in fact most human genetic variation is in the form of CNVs and other SVs.

CNVs are thought to occur predominately from three methods: non-allelic homologous recombination (NAHR), non-homologous end-joining (NHEJ), and fork stalling and template switching (FoSTeS), as reviewed in [26] and illustrated in Figure 1.1. Briefly, NAHR is similar to normal homologous allelic recombination except that the Holliday junction forms between two homologous elements that are not alleles. The non-allelic homologous elements are usually repeats. NHEJ occurs when cellular DNA undergoes a double-strand break (DSB) that is then repaired by end-joining and ligation with another cleaved DNA end. DSBs and NHEJ are integral for V(D)J recombination that leads to immune variability in higher eukaryotes including humans. However, it can also lead non-physiological outcomes that are associated with disease. The FoSTeS model for CNV formation refers to the periodic disassociation of the polymerase-nascent strand complex from the replication fork and its re-association with a downstream replication fork (causing a deletion) or an upstream replication fork (causing an insertion). This is a consequence of the presence of multiple simultaneous replication forks in a replicating genome.

1.3.4 Methods for SV Detection

Detecting SVs poses an obvious challenge to current high-throughput sequencing technologies that rely on short read lengths that are computationally mapped to a

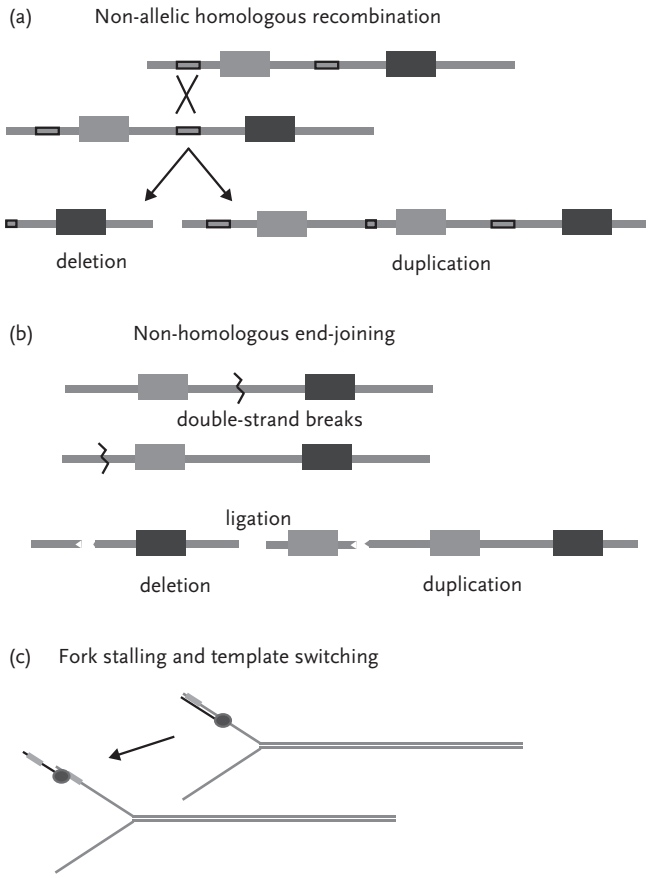

Figure 1.1 Schematic of major mechanisms of CNV formation. (a) Non-allelic homologous recombination (NAHR). Briefly, NAHR occurs when non-allelic homologous regions (i.e., genomic repeats of high enough sequence similarity) recombine during genome replication. This facilitates CNV creation as depicted in the schematic. (b) Non-homologous end-joining (NHEJ). NHEJ occurs when a double-strand break (DSB) is repaired by joining a DNA break-point with a break-point of a different strand, not the original strand where the break occurred. (c) Fork stalling and template switching (FoSTeS). FoSTeS occurs when, during genome replication the DNA polymerase at a replication fork switches to another replication fork along with the nascent strand, thereby either skipping template DNA stretches (causing deletions) or repeating them (causing duplications).

reference genome to re-sequence a human genome. Long repeats and rearrangements will not be distinguished from each other purely based on sequence when short read lengths are used to compile the full sequence. Thus, several methods have been developed to overcome this problem (summarized in Figure 1.2).

Comparative genomic hybridization (array-CGH) uses an array with oligonucleotides corresponding to the DNA sequence being probed to test a sample genome as

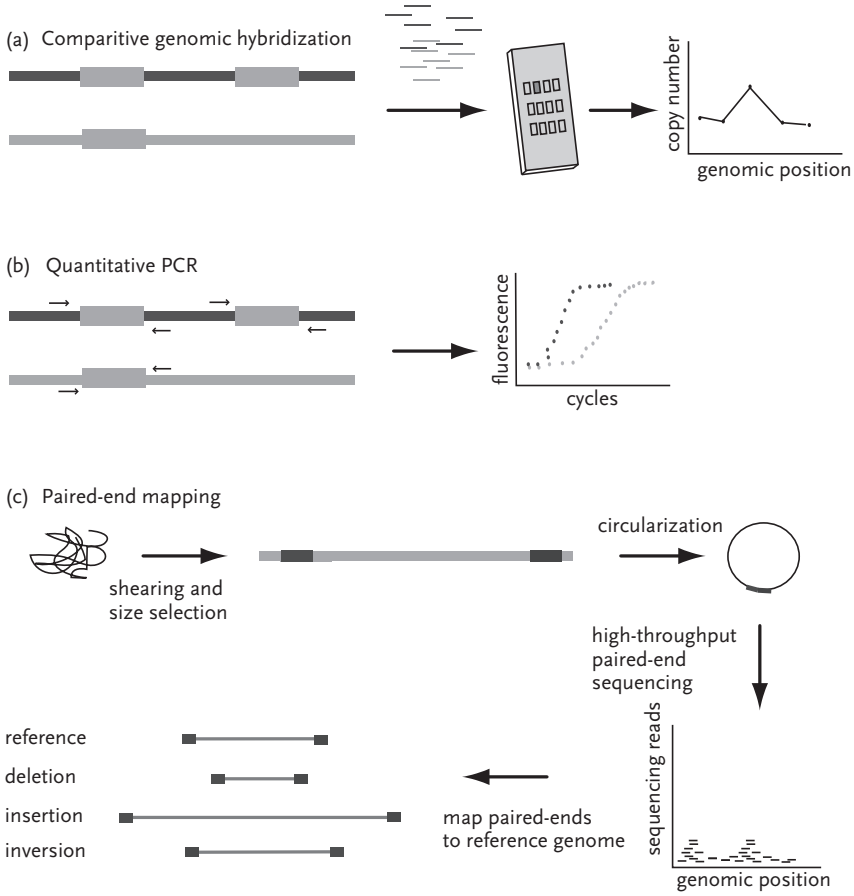

Figure 1.2 Methods for structural variant detection. (a) Comparative genomic hybridization (CGH). CGH involves fragmenting a reference DNA sample and the test sample that are coated with a green fluorophore for one and red for the other. The samples are then hybridized on an array and different copy numbers for each of the fragmented segments are reported as unequal mixtures of the two fluorophores. (b) Quantitative polymerase chain reaction (qPCR). In qPCR primers are used to amplify a desired genomic region. Various kinds of fluorescent probes are used to detect the amount of double stranded DNA synthesized in the reaction. Thus, one is able to quantify the amount of product and hence template, which can be used to detect CNVs. (c) Paired-end mapping (PEM). In PEM, paired-end high-throughput sequencing is used to map SVs at a resolution of up to 3 kb. First, genomic DNA is sheared. Then, only 3-kb fragments are selected. These are circularized and the circular DNA is then fragmented and sequenced using a high-throughput sequencing platform. The sequence of the junction from the circularized DNA provides information as to the end points of the 3-kb fragment. These end-point sequences are then mapped to the reference genome. The distance between the end-points in the reference genome is compared to the expected 3-kb distance (found in the test sample genome). Those fragments that are longer than 3 kb in the reference genome provide evidence of a deletion in the test sample, those that are shorter imply an insertion, and those that are rearranged imply an inversion. PEM is not the only method that takes advantage of high-throughput sequencing to map SVs. Read depth analysis detects CNVs by finding candidate regions to which reads map at higher frequencies in a sample genome as compared to a reference genome.

compared to a reference control genome for differences in hybridization to the array [21, 27]. The DNA from each sample is usually labeled with different fluorophores and the binding intensity is thus assessed. The genome with the higher copy number of a particular region hybridizes more intensely to the oligo probe [21].

Quantitative PCR and variations of it, like molecular copy-number counting (MCC), are also used to detect and quantify CNV [21]. MCC involves diluting the DNA samples to such a high degree that there are only a small number of molecules per well in a 96-well plate. Then, multiplex, single-molecule PCR is used to detect the presence of desired loci, thus quantifying the number of copies of the loci [21]. The most straightforward method for CNV detection is deep sequencing, where the occurrence of reads computationally mapped to the genome is used as an estimate for the frequency of occurrence of the sequence to which the reads map. In other words, if a gene occurs five times as frequently in genome A as it does in genome B, it is likely that there will be five times as many reads mapped to the sequence of the locus in genome A as in genome B.

The above-mentioned methods fall short in identifying CNVs at high resolution, and more generally, do not accurately identify the genomic locations of SVs. Recent work in our laboratory addressed the challenge of how to use high-throughput sequencing to identify SVs as small as 3 kb and their genomic breakpoints at nucleotide level resolution [18]. We developed a method called paired-end mapping (PEM) where sample genomes are fragmented and 3-kb fragments are selected and sequenced with the 454 sequencing platform [18]. The paired ends are then computationally mapped to the reference genome and the distances between the ends are compared to the expected 3-kb size [18]. Deletions in the sample genome mean that the reference genome distance between the paired ends is greater than 3-kb, whereas the presence of insertions in the sample genome means that the reference genome distance is smaller than 3-kb. Similarly, inversions have different breakpoints in the reference as compared to the sample [18]. We applied this method to construct the first high-resolution map of SVs in the human genome by comparing an African individual and a European individual to the reference genome [18]. More than 1000 SVs were found using this method, 30% of which were less than 5 kb.

1.3.5
Transposons and Retrotransposons

Much of the genome is composed of repetitive elements that have been inserted (and continue to be copied or moved and inserted if they are active) into positions on the genome throughout evolutionary history. Approximately 40% of the human genome is thought to be composed of such elements [19, 28–30]. They are thought to be remnants of viruses or other reverse transcribed elements. Mobile elements or their remnants are called transposons if they are mobile DNA segments that are spliced into different genomic positions with the help of facilitators like transposase. Retrotransposons are first transcribed into RNA and then retrotranscribed into DNA that is then inserted into the genome.

Retrotransposons make up approximately one-third of the human genome [19, 30]. Two predominant subclasses of retrotransposons are long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LINEs are the more predominant element, composing approximately 20% of the human genome [19]. LINEs encode a reverse transcriptase that reverse transcribed the RNA that is transcribed from the LINE DNA by RNA Pol II. SINEs do not encode a reverse transcriptase and are likely a result of reverse transcribed small cellular RNAs such as miRNAs, snoRNAs, tRNAs, and others.

1.4 Regulatory Information

Though only about 2% of the human genome codes for proteins, much of the rest of the DNA is highly conserved with other species and among individuals [19, 31–33]. This has led many to speculate that much of the genome has a functional role in human biology through the regulation of the quantity and timing of protein expression. In fact, King and Wilson postulated in 1975 that gene regulation may be the main driver of phenotypic divergence in evolution [34]. Numerous functional genomics studies have suggested that non-coding DNA is largely responsible for gene expression regulation through interaction with transcriptional regulatory machinery [35–41]. Regulatory genomic elements are discussed in this section and schematically summarized in Figure 1.3a.

To better understand the genomic elements that regulate gene expression, researchers over the past 20 years have investigated variation in gene expression in different individuals, species, and tissues in variable experimental conditions. Already there is evidence that alleles with regulatory sequence mutations are responsible for susceptibility to human diseases including autoimmune, psychiatric, neoplastic, and neurodegenerative diseases [42]. As reviewed previously, mutations in *cis*-regulatory sequences of specific genes that have been linked to a variety of conditions, for example: AVPR1A regulation is linked to creative dance performance in humans and parental care in rodents, HTR2A is linked to obsessive compulsive behavior, and MMP3 is linked to the risk of heart disease [43]. A 58-kb locus on chromosome 9 has been linked to an increased risk for coronary artery disease, though the nearest protein-coding gene is located more than 60 kb away [44]. Glucocorticoid-remedial aldosteronism (GRA) is caused by the fusion of a regulatory region of one gene with a different gene, thereby perturbing the gene regulation necessary for proper expression of aldosterone synthase at the right time in the right tissues [17].

1.4.1 Classes of Regulatory Elements

There are two approaches to classifying regulatory elements in the genome: by their mechanism of affecting gene expression and by their position relative to the gene(s) they control. The mechanism of action distinction gives us two types of genetic

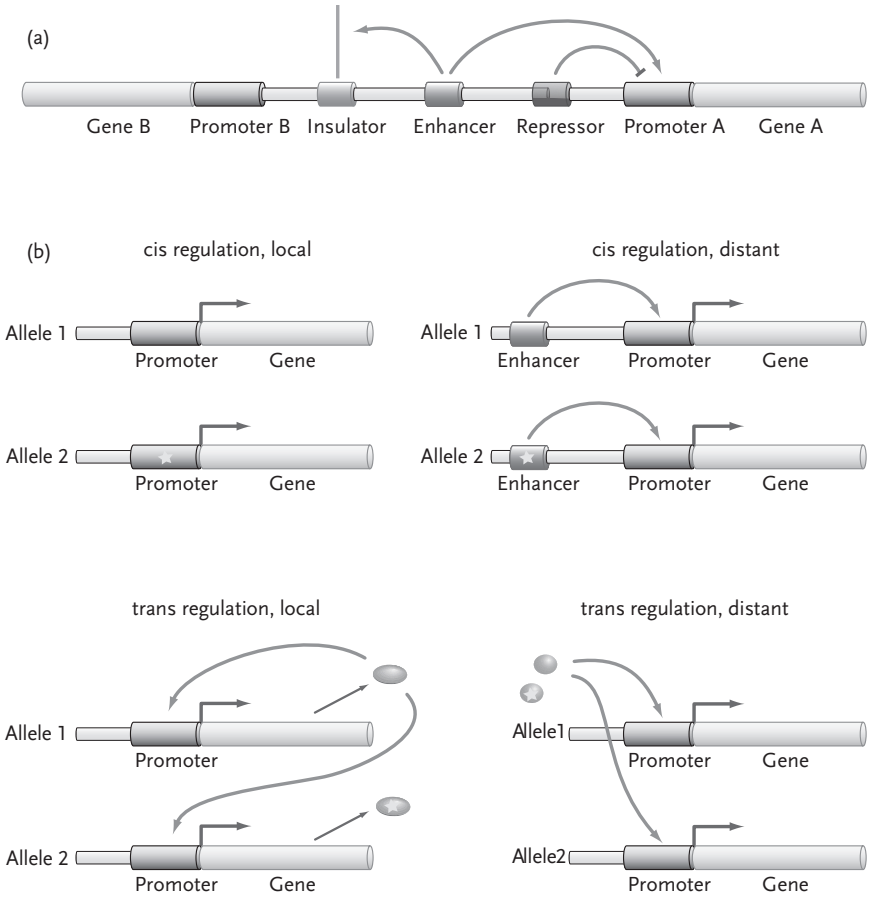

Figure 1.3 Regulatory elements. (a) Classes of regulatory elements. Promoter sequences bind the general transcription machinery, including transcription factors (TFs) and RNA Pol II. As depicted, enhancer and repressor sequences positively or negatively regulate transcription via interactions with the promoter sequence, usually involving TF binding to the enhancer region and cofactors facilitating the interaction with the promoter. In addition insulator elements prevent enhancers from activating other genes. (b) Cis/trans and local/distant regulation. The star indicates the regulatory variant. The following four scenarios can be distinguished. (i) Promoters where TFs bind represent local cis regulation. A mutation in the promoter on one allele will only affect gene expression on that allele, not on the other allele. (ii) Cis regulation can also occur through distant elements like enhancers. (iii) Trans regulation can occur through local elements like in the depicted example of an auto-regulatory system. If the gene product is itself involved in regulating the gene then a mutation in the gene product will affect both alleles. (iv) Distant trans-acting regulators are elements like TFs that are expressed in an alternate location on the genome but affect both alleles under the control of the promoter that they bind.

control: *cis* and *trans* (Figure 1.3b). *Cis*-regulatory sequences control genes on the allele on which they are present. In contrast, *trans*-regulatory control is allele-independent. Variation in a *trans*-regulator will affect both alleles under its control. A common example of *cis* regulation is the transcription factor (TF) binding sequences in the promoter regions of genes. Variation in these sequences can cause changes in the way a TF binds and regulates gene expression, but only on the allele that has the sequence changes. The TF itself acts in *trans*. Changes in the sequence or the expression of the TF affect both of the alleles that it regulates. Similarly, variation in the 3′ UTR of mRNAs affecting miRNA binding or mRNA stability is *cis*-regulatory, whereas changes in the expression or sequence of the miRNA are *trans*.

Local regulatory elements are those located near the gene whose expression they affect. Distant regulators are located far from the gene(s). Local regulatory elements include TF binding sites, miRNA binding sites, and splicing variants. Distant regulators include enhancer, repressor, and insulator sequences, transcription factors, and miRNAs. The definitions of local and distant are of course relative, and much is still unknown about the 3-D structure of chromatin in the nucleus and how it relates to gene expression [44]. Furthermore, location distinctions tend to be conflated with the mechanistic *cis* and *trans* classification because *cis* regulation tends to occur through local regulators, such as TF binding sites, and *trans* through distant ones, such as TFs. However, in principle, the two classification categories are independent. Indeed, there are counter examples to the above-mentioned trends: local mutations that disrupt the function of an autoregulator act in *trans*, distant enhancers can act both in *trans* and *cis*, and mutations influencing regional chromatin structure can affect the regulation of a large number of genes and thus can be distant but act in *cis* [45]. Nonetheless there is abundant evidence that regulatory sequences nearer to the gene they control tend be *cis*: Ronald *et al.* estimate that in yeast 75% of local expression quantitative trait loci (eQTLs; measurements of gene expression treated as a quantitative trait and then associated with a particular locus) act in *cis* and the rest in *trans* [46]. Similarly, a report by Pickrell *et al.* of deep sequencing of the human lymphoblastoid cell line (LCL) transcriptome revealed that only 5% of *cis* regulators are greater than 200 kb away from the start of the gene that they influence [47].

1.4.2
Transcription Factor Binding Motifs

Cis-regulatory elements are often promoter sequences proximal to genes and are bound by TFs that then recruit RNA Polymerase. Because of the importance of TF binding to gene expression, much work has gone into mapping TF binding sites in both the yeast and human genomes under different conditions or in different tissues (as well as in other model systems). The hope is that understanding the sequence determinants of TF binding will help bridge the knowledge gap between sequence and function. An important method for probing TF binding is chromatin immunoprecipitation (ChIP). A key step in ChIP involves a technique first developed by Alexander Varshavsky and colleagues where formaldehyde is used to

cross-link proteins to DNA, thereby capturing the state of protein–DNA interaction in the cell [48]. The protein–DNA complex is then purified using antibodies for the specific protein in question. Early work in mapping TF binding relied on identifying the protein-bound DNA by microarray hybridization following ChIP (ChIP-chip) [38, 49]. Since the advent of high-throughput sequencing this method has been modified, replacing array hybridization with deep sequencing of the purified DNA, resulting in ChIP-Seq [36, 50]. The sequence is then mapped to the genome to ascertain the genomic position of the DNA–protein interaction.

TF binding to promoters of orthologous genes is significantly divergent among yeast species and between human and mouse [35, 37, 51]. TF binding divergence far exceeds the sequence divergence of the genes themselves. This may help partially explain the large phenotypic differences between species; divergent TF binding could be evidence of divergent gene regulation. However, in many cases, TF binding divergence between species does not seem to correlate well with expression divergence [52]. When binding site sequences in *S. paradoxus* and *S. mikatae* were mutated to the *S. cerevisiae* orthologous promoter sequence, gene expression changed in only three out of the 11 genes tested [53]. Tirosh *et al.* directly compared the divergence of the mating response TF, STE12, binding motif in promoter sequences of *S. cerevisiae*, *S. paradoxus*, and *S. mikatae* and the expression divergence of mating genes under the regulation of STE12 in these species [52]. The results suggest that approximately half of the expression divergence in the mating pathway among the three yeast species can be explained by the loss or gain of STE12 binding motifs. The remaining half of the expression divergence may be due to epigenetic inherited elements, complicated combinatorial regulation that confounds the correlation between STE12 binding and expression, or a yet unknown mechanism that creates a distinction between actively bound TFs and bound but not active TFs.

Only recently have we begun to appreciate the complexity and the importance of more distant regulation sequences such as enhancers, repressors, and insulators [44]. Additional factors may be involved in guiding TF binding and activating bound TFs. TFs can be repressors, but also one can imagine a scenario where TF binding in one region does not affect gene expression because its purpose is to titrate out TF molecules so that they do not bind somewhere else [54]. This model, reviewed by Segal and Widom, describes transcriptional regulation as a dynamic system of competition for DNA binding among different TFs, nucleosomes, and potentially other factors (for competitive binding of TFs to nucleosomes, see Chapter 5). The transcription level is then determined by the equilibrium occupancy of binding sites of all these factors according to their concentration and inherent DNA binding affinity. However, on the genome-wide scale this model does not entirely explain how TFs bind to specific promoters. TFs have preferences for specific sequence motifs, usually 5–10 bp in length, which can be expressed as position weight matrices that describe the importance of base pairs for binding at each position [55]. However, a simple calculation shows that in the human genome there are likely to be about 10^6 sequences that satisfy a particular TFs binding motif: if the motif is 5 bp, a random sequence of 3×10^9 bp (the haploid length of the human genome) is likely to contain $1/4^5 \times 3 \times 10^9 \approx 3 \times 10^6$ occurrences

of the specific 5 bp sequence [54]. It would take considerable evolutionary energy to counter these statistics to select against so many occurrences of such a short sequence fragment. Based on the quantity of TF proteins produced in the cell, this means that the ratio of TF to potential binding sequence ranges from 1 : 10 to 1 : 1000 [54]. This problem is at least partially addressed by the cell labeling entire regions of the genome as off limits to TFs via chromatin states in which the strength of the histone octamer DNA interaction in the nucleosome precludes TF binding (Chapters 3 and 5) [36, 54, 56]. Perhaps there are further layers of partitioning and compartmentalization of functional genomic elements that cannot be dissected from studying sequence binding biochemistry alone.

Nonetheless, there are striking examples of TF binding motif gain/loss that correlate to large-scale rewiring of gene regulation with stark phenotypic affects. The expression of cytoplasmic ribosomal proteins (RP) and mitochondrial ribosomal proteins (MRP) is highly correlated in *Candida albicans* yeast, but not correlated in *S. cerevisiae*, where instead MRP genes co-express with stress-response genes [57]. This is consistent with the physiology of the two organisms; *S. cerevisiae* have the ability to grow rapidly in anaerobic conditions whereas *C. albicans* can only grow in the presence of oxygen. MRP and RP genes in *C. albicans* are enriched in the promoter sequence with the *cis*-regulatory sequence "AATTTT." In *S. cerevisiae* this motif is present in RP genes but is missing in MRP genes [57]. It seems as though decoupling of the MRP and RP gene expression programs was a necessary step in the evolution of rapid anaerobic growth, and there is evidence to suggest that the wholesale loss of a *cis*-regulatory element in a group of genes contributed to this reprogramming. This could have occurred during the whole genome duplication event in the evolutionary history of yeast, where large-scale duplication allowed for the neofunctionalization of genes and regulatory elements [58].

1.4.3 Allele-Specific Expression

Elucidating the regulatory mechanism involved behind eQTLs usually involves a method for distinguishing whether gene expression is allele-specific or not, based on the sequence of the transcribed alleles. The eukaryotic genome usually has two alleles of each gene. To determine which allele is being transcribed one can identify any sequence heterogeneity in the exonic portion of the gene and quantitatively sequence the mRNA or hybridize to arrays (where the array hybridization is sensitive to the sequence differences between the alleles). Then, each transcribed RNA is mapped to one of the alleles based on its sequence polymorphisms. An imbalance in the ratio of transcripts from the two alleles points to allele-specific expression.

Recently, Tirosh *et al.* constructed an elegant system to measure the relative contribution of *cis* QTLs and *trans* QTLs to genome-wide expression divergence in related yeast species [59]. The authors used microarrays to measure gene expression in a hybrid of *S. cerevisiae* and *S. paradoxus*. Thus, for each gene one allele comes from one species and the other allele from the other species. Sequence

polymorphisms that correspond to species-specific gene expression can be classified as *cis* and those variations that affect expression from both alleles are *trans*. They find that *cis* sequence variation accounts for most of the expression differences between the two species, but that *trans* variants allow for differential response to changing environmental conditions. Approximately two-thirds of the *trans* regulatory differences identified exerted a differential effect on gene expression in response to particular experimental conditions. By contrast, *cis* sequences were condition-independent: 77% influenced gene expression regardless of the conditions in which the yeast was grown. This is consistent with our understand of gene regulation in that tweaking a small number of *trans* regulators like TFs can trigger a response involving the expression of many genes, so evolution in *trans* regulators is the likely organismal adaptation to changing environments. In contrast, *cis* polymorphisms tend to be binding sites that can vary between alleles and would need to be mutated in a large number of promoters to affect a large phenotypic response, so evolution in *cis*-regulatory regions may be a less effective way of dealing with environmental change, barring the case of massive genomic duplication or loss.

Allele-specific expression has also been measured in human systems, taking advantage of the growing database of SNPs to identify heterozygous polymorphisms in transcripts that serve as allele markers. Several technologies have been used to genotype transcripts: RNA-Seq [47, 60, 61], "padlock" probe capture of DNA regions and deep sequencing (RNA allelotyping [62, 63], padlock method [64]), and array-based SNP genotyping [65]. In each of the cases mRNA is identified by the 3′ poly-A tail and then reverse transcribed into cDNA which is then sequenced or hybridized on an array. Most work until now has been done in LCLs from the HapMap study, though some studies have looked at fibroblasts and differentiated iPS cells [62] and embryonic stem cells [63]. These reports, from different groups and different technologies, present data suggesting that allele-specific expression is in fact widespread in the human genome. Zhang *et al.* [63] show that 11–22% of the heterozygous SNPs that they found in cell lines from two human subjects are expressed at an imbalance of greater than 0.1, that is, the ratio between the alleles is >0.6 or <0.4. Less conservatively, Ge *et al.* [65] call anything that has an imbalance of greater than 0.5 an instance of allele-specific expression. They report that approximately 30% of human genes have differential expression of alleles. Interestingly, $\sim$5.8% of SNPs in the HapMap database link to at least one genomic region where allele-specific expression occurs [65]. This suggests that common SNPs may play a large role in allele-specific expression through *cis* regulation of transcription and supports the hypothesis that single nucleotide differences between human individuals can have significant phenotypic effects, even if they are located in non-coding genomic regions. In fact, there is a significant overlap between SNPs linked to allele-specific expression and common SNPs associated with autoimmune disease in genome-wide association studies (GWAS) for type I diabetes, Crohn's disease, and systemic lupus erythematosus [65].

There is also evidence that allelic expression is tissue-specific. Alleles express at different ratios in lymphoblastoids, primary fibroblasts, primary keratinocytes, and

embryonic and iPS stem cells [62, 63]. The differential effect of *cis*-regulatory sequences has also been observed in adipose and blood tissue [66]. In a cohort GWAS of obesity, gene expression in adipose tissue correlated to the obesity phenotypes observed in the human subjects, whereas gene expression profiles from the blood tissue did not [66]. This highlights the complexity of genetic regulation: regulatory sequence changes on alleles may only be detectable in certain conditions or tissues or may be masked or enhanced by changes in *trans* factors elsewhere in the genome.

1.5 Individual Genetic Polymorphisms and Their Effect on Gene Expression

Recent work in our laboratory has helped address how these individual variations may affect TF binding and gene expression. Binding of the TF STE12 was assessed by ChIP-Seq in related strains of *S. cerevisiae* yeast [67]. Binding variation correlated to changes in expression in approximately one-third of the cases. Significantly, much of the binding variation could be explained by SNPs and insertions/deletions in the binding motif of STE12, demonstrating on the molecular level that intra-species variation affects TF binding and often gene expression [67].

A similar question was asked for inter-individual variation in humans. Nine LCLs from individuals enrolled in the HapMap study, one LCL from the genome-wide SV mapping study [68] and one LCL from a chimpanzee were tested for binding of RNA Polymerase II and NFκB, an immune response regulator [69]. A comparison of the distribution of binding peaks shows that binding between two individuals varies by an average of 7.5 and 25.0% for NFκB and RNA Pol II, respectively. Interestingly, binding sites further than 1 kb away from the transcriptional start site of a gene exhibited greater variance than more local binding sites [69]. A total of 35 and 26% of the binding difference for NFκB and RNA Pol II, respectively, between individuals was correlated to the occurrence of SNPs or SVs within their binding regions. Kasowski, Grubert *et al.* termed these genetic variations binding-SNPs (B-SNPs) and binding-SVs (B-SVs) because they influence TF binding [69]. Binding differences between individuals correlated significantly to gene expression differences, as measured by RNA-Seq.

1.6 Conclusion

We have come a long way in our understanding of the functional elements of the genome. In addition to the protein code, new technologies allow us to begin to appreciate the complexities of the non-coding sequence. Genetic mapping and allele-specific expression approaches combined with high-throughput sequencing

allow us to better understand the nature of expression variation, and ChIP lets us explore the molecular basis of this variation more deeply. In so doing these technologies help us to understand the genetic and molecular basis of human variation and disease.

Acknowledgments

We are grateful to Rajini Haraksingh, Hogune Lm, Daniel Kaganovich, and Alex Pollen for helpful discussions and comments on the manuscript.

References

1 Hong, X., Scofield, D.G., and Lynch, M. (2006) Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol*, **23**, 2392–2404.

2 Rozowsky, J., Wu, J., Lian, Z., Nagalakshmi, U., Korbel, J.O., Kapranov, P., Zheng, D., Dyke, S., Newburger, P., Miller, P., and Gingeras, T.R. *et al.* (2006) Novel transcribed regions in the human genome. *Cold Spring Harb Symp Quant Biol*, **71**, 111–116.

3 Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., and Schmidt, D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

4 Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *11206552*, **456**, 470–476.

5 Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *1090005*, 47.

6 Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., and Cabili, M.N. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *11206552*, **458**, 223–227.

7 Wu, J.Q., Du, J., Rozowsky, J., Zhang, Z., Urban, A.E., Euskirchen, G., Weissman, S., Gerstein, M., and Snyder, M. (2008) Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biol*, **9**, R3.

8 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, **5**, 621–628.

9 Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *11206552*, **457**, 1038–1042.

10 Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *11206552*, **457**, 1033–1037.

11 Brodsky, A.S., Meyer, C.A., Swinburne, I.A., Hall, G., Keenan, B.J., Liu, X.S., Fox, E.A., and Silver, P.A. (2005) Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol*, **6**, R64.

12 Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G.,

Chepelev, I., and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *17604720*, **129**, 823–837.

13 Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA*, **106**, 7507–7512.

14 Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., and Milos, P.M. (2009) Direct RNA sequencing. *11206552*, **461**, 814–818.

15 Zhang, Z. and Gerstein, M. (2004) Large-scale analysis of pseudogenes in the human genome. *Curr Opin Genet Dev*, **14**, 328–335.

16 Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S.W., Lu, Y., Denoeud, F., Antonarakis, S.E., Snyder, M., Ruan, Y. *et al.* (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res*, **17**, 839–851.

17 Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet*, **14**, 417–422.

18 Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., and Taillon, B.E. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

19 Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., and Gerstein, M.B. (2010) Annotating non-coding regions of the genome. *Nat Rev Genet*, **11**, 559–571.

20 Schlotterer, C. and Tautz, D. (1992) Slippage synthesis of simple sequence DNA. *Nucl Acids Res*, **20**, 211–215.

21 Dear, P.H. (2009) Copy-number variation: the end of the human genome? *Trends Biotechnol*, **27**, 448–454.

22 Korbel, J.O., Kim, P.M., Chen, X., Urban, A.E., Weissman, S., Snyder, M., and Gerstein, M.B. (2008) The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol*, **18**, 366–374.

23 McCarroll, S.A., Kuruvilla, F.G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., and Elliott, A.L. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, **40**, 1166–1174.

24 Bailey, J.A. and Eichler, E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, **7**, 552–564.

25 Conrad, D., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T., Barnes, C., Campbell, P., and Fitzgerald, T. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *11206552*, **464**, 704–712.

26 Gu, W., Zhang, F., and Lupski, J.R. (2008) Mechanisms for human genomic rearrangements. *Pathogenetics*, **1**, 4.

27 Hurles, M., Dermitzakis, E., and Tyler-Smith, C. (2008) The functional impact of structural variation in humans. *11418218*, **24**, 238–245.

28 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., and Gocayne, J.D. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

29 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., and Funke, R. *et al.* (2001) Initial sequencing and analysis of the human genome. *11206552*, **409**, 860–921.

30 Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010) LINE-1 retrotransposition activity in human genomes. *17604720*, **141**, 1159–1170.

31 Elgar, G. and Vavouri, T. (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *11418218*, **24**, 344.

32 Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004) Ultraconserved elements in the human genome. *1090005*, **304**, 1321.

33 Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., and Rosenbloom, K.R. *et al.* (2006) Forces shaping the fastest evolving regions in the human genome. *16121257*, **2**, e168.

34 King, M. and Wilson, A. (1975) Evolution at two levels in humans and chimpanzees. *1090005*, **188**, 107–116.

35 Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007) Divergence of transcription factor binding sites across related yeast species. *1090005*, **317**, 815.

36 Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R.P., and Lee, W. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *11206552*, **448**, 553.

37 Odom, D.T., Dowell, R.D., Jacobsen, E. S., Gordon, W., Danford, T.W., Macisaac, K.D., Rolfe, P.A., Conboy, C. M., Gifford, D.K., and Fraenkel, E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *16783381*, **39**, 730–732.

38 Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., and Volkert, T.L. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *1090005*, **290**, 2306.

39 Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H. Y. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *17604720*, **129**, 1311.

40 Vastenhouw, N.L., Zhang, Y., Woods, I. G., Imam, F., Regev, A., Liu, X.S., Rinn, J., and Schier, A.F. (2010) Chromatin signature of embryonic pluripotency is established during genome activation. *11206552*, **464**, 922–926.

41 Visel, A., Blow, M., Li, Z., Zhang, T., Akiyama, J., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., and Afzal, V. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *11206552*, **457**, 854–858.

42 Skelly, D., Ronald, J., and Akey, J. (2009) Inherited variation in gene expression. *19630563*, **10**, 313–332.

43 Wray, G.A. (2007) The evolutionary significance of cis-regulatory mutations. *17047685*, **8**, 206.

44 Visel, A., Rubin, E., and Pennacchio, L. (2009) Genomic views of distant-acting enhancers. *11206552*, **461**, 199–205.

45 Rockman, M. and Kruglyak, L. (2006) Genetics of global gene expression. *17047685*, **7**, 862.

46 Ronald, J., Brem, R., Whittle, J., and Kruglyak, L. (2005) Local regulatory variation in *Saccharomyces cerevisiae*. *16121257*, **1**, e25.

47 Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J., Stephens, M., Gilad, Y., and Pritchard, J. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *11206552*, 520.

48 Solomon, M., Larsen, P., and Varshavsky, A. (1988) Mapping protein-DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *17604720*, **53**, 937–947.

49 Iyer, V., Horak, C., Scafe, C., Botstein, D., Snyder, M., and Brown, P. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *11206552*, **409**, 533–538.

50 Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *17047685*, **10**, 669–680.

51 Schmidt, D., Wilson, M., Ballester, B., Schwalie, P., Brown, G., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C., Mackay, S., and Talianidis, I. *et al.* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of

transcription factor binding. *1090005*, **328**, 1036–1040.

52 Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. (2008) On the relation between promoter divergence and gene expression evolution. *18197176*, **4**, 159.

53 Doniger, S.W. and Fay, J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. *17530920*, 3.

54 Segal, E. and Widom, J. (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet*, **10**, 443–456.

55 Harbison, C., Gordon, D., Lee, T., Rinaldi, N., Macisaac, K., Danford, T., Hannett, N., Tagne, J., Reynolds, D., Yoo, J., and Jennings, E. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *11206552*, **431**, 99.

56 Tirosh, I. and Barkai, N. (2008) Two strategies for gene regulation by promoter nucleosomes. *18448704*, 4.

57 Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J., and Barkai, N. (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *1090005*, **309**, 938.

58 Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007) Natural history and evolutionary principles of gene duplication in fungi. *11206552*, **449**, 54–61.

59 Tirosh, I., Reikhav, S., Levy, A., and Barkai, N. (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *1090005*, **324**, 659–662.

60 Montgomery, S.B. and Dermitzakis, E.T. (2009) The resolution of the genetics of gene expression. *19808798*, **18**, R211–215.

61 Montgomery, S., Sammeth, M., Gutierrez-Arcelus, M., Lach, R., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *11206552*, **464**, 773–777.

62 Lee, J.H., Park, I.H., Gao, Y., Li, J.B., Li, Z., Daley, G.Q., Zhang, K., and Church, G.M. (2009) A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet*, **5**, e1000718.

63 Zhang, K., Li, J., Gao, Y., Egli, D., Xie, B., Deng, J., Li, Z., Lee, J., Aach, J., Leproust, E., and Eggan, K. *et al.* (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *19349980*, **6**, 613–618.

64 Li, J.B., Levanon, E.Y., Yoon, J.K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *1090005*, **324**, 1210–1213.

65 Ge, B., Pokholok, D.K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D.J., Le, J., Koka, V., Lam, K.C., GagnÈ, V., and Dias, J. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *16783381*, **41**, 1216–1222.

66 Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., and Mouy, M. *et al.* (2008) Genetics of gene expression and its effect on disease. *11206552*, **452**, 423.

67 Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M., and Snyder, M. (2010) Genetic analysis of variation in transcription factor binding in yeast. *11206552*, **464**, 1187–1191.

68 Korbel, J., Urban, A., Affourtit, J., Godwin, B., Grubert, F., Simons, J., Kim, P., Palejev, D., Carriero, N., Du, L., and Taillon, B. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *1090005*, **318**, 420.

69 Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., and Hong, M.Y. *et al.* (2010) Variation in transcription factor binding among humans. *1090005*, **328**, 232–235.