Annotating non-coding regions of the genome

*Roger P. Alexander**‡, Gang Fang**‡, Joel Rozowsky‡, Michael Snyder§ and Mark B. Gerstein**‡||*

Abstract | Most of the human genome consists of non-protein-coding DNA. Recently, progress has been made in annotating these non-coding regions through the interpretation of functional genomics experiments and comparative sequence analysis. One can conceptualize functional genomics analysis as involving a sequence of steps: turning the output of an experiment into a 'signal' at each base pair of the genome; smoothing this signal and segmenting it into small blocks of initial annotation; and then clustering these small blocks into larger derived annotations and networks. Finally, one can relate functional genomics annotations to conserved units and measures of conservation derived from comparative sequence analysis.

When it was realized in the late 1960s[1] how much of the genome does not code for proteins, non-coding regions were thought to be non-functional and were labelled as 'junk DNA'[2]. When discussions began about sequencing the human genome, there was vigorous debate about whether to avoid repeat regions and focus only on protein-coding regions. Some thought that sequencing 'junk DNA' would be a waste of time and money[3,4] — a question we face again with the advent of targeted exome sequencing[5–7]. In the era of personal genomics, is it important to sequence whole human genomes, or can we focus only on protein-coding exons, which comprise less than 2% of the genome sequence[8,9]?

One way to address this question is to consider the functions attributed to the non-coding genome. In particular, a growing number of non-coding transcripts have been assigned roles in gene regulation and RNA processing[10]. Cross-species sequence comparisons have identified conserved non-coding elements (NCEs) that are candidates for function[11–13]. SNPs in many non-coding regions have been linked to disease in genome-wide association (GWA) studies[14–16], as might be expected from the preponderance of non-coding SNPs assayed on genotyping chips. Studies of structural variants in the genome have shown that many large blocks in non-coding regions vary among individuals, and some of these structural variants have been linked to disease[17,18]. Copy-number variants that contain genes often have their origin in recombination between non-coding repeat regions[17,19]. In addition, non-coding DNA provides a historical record of genome evolution, as it contains 'fossils' of molecules that were historically active. These fossils, which are often well-known repetitive elements, can mediate genome evolution itself, causing mistakes in recombination that lead to duplication or the deletion of functional sequences.

Where do we stand in the effort to annotate the non-coding genome? In part to address the role of non-coding DNA, in 2005 the National Human Genome Research Institute (NHGRI) launched the Encyclopedia of DNA Elements (ENCODE) Project with the aim of annotating all elements in the human genome. A pilot phase examining a representative 1% of the genome with a wide array of functional genomics experiments was completed in 2007 (REF. 20). The scope of ENCODE has now expanded to the whole-genome scale, and the NHGRI has launched a parallel project, modENCODE, to annotate the genomes of the model organisms *Drosophila melanogaster* and *Caenorhabditis elegans*[21] and to relate these annotations to the human genome.

How should we think about annotating the non-coding genome? As an analogy, consider how we might annotate a written document[22–24]. A first step would be to notate words, phrases and longer blocks repeated at different levels of similarity in the document. Next, we might highlight certain functional elements of the document, such as the title, block quotes or subheadings, in a distinct font. Finally, we might try to interrelate the functions of these highlighted elements to the patterning of repeated words and phrases.

Targeted exome sequencing
A technique that involves filtering genomic DNA by capturing regions of interest (often protein-coding exons) on a microarray, then sequencing the captured DNA using next-generation techniques.

**Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.*
‡Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
§Department of Genetics, Stanford University, Stanford, California 94305, USA.
||Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA.
Correspondence to M.B.G.
e-mail:
mark.gerstein@yale.edu
doi:10.1038/nrg2814

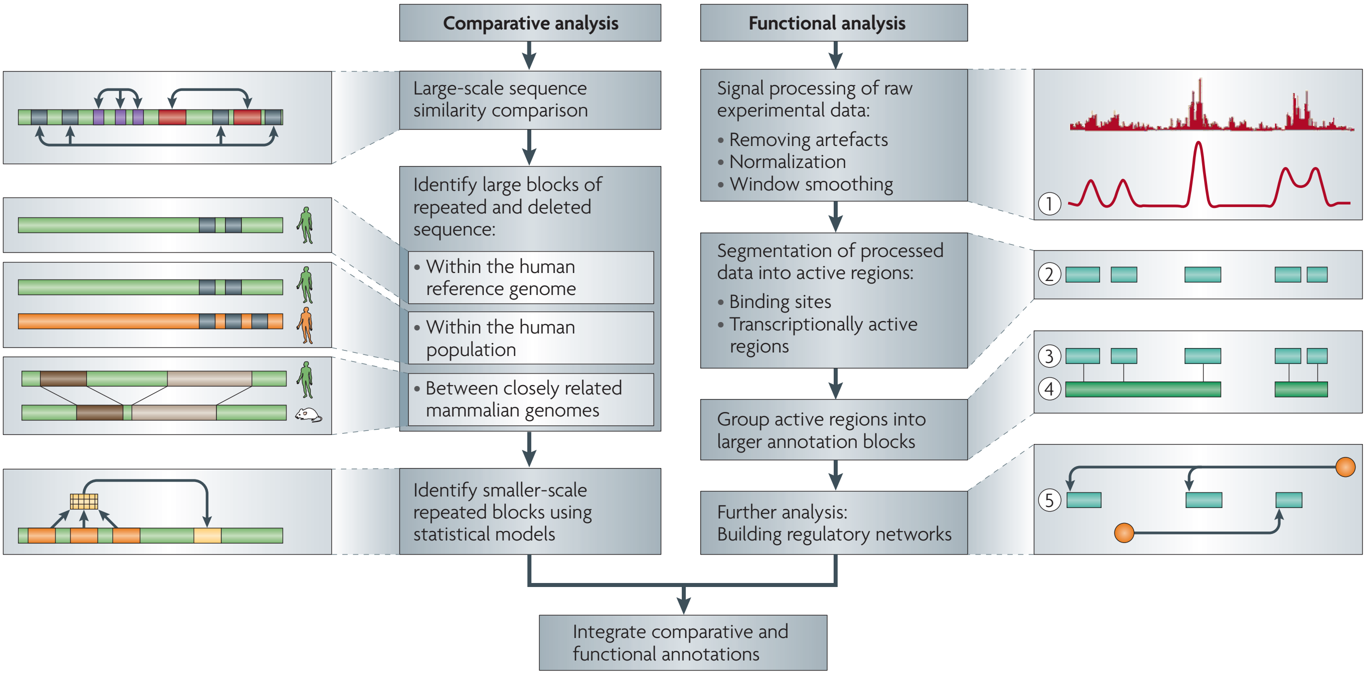


Figure 1 | **Annotation process for non-coding regions: an overview.** The annotation process includes two parallel pipelines for comparative sequence analysis (comparative analysis) and functional genomics analysis (functional analysis) of experimental data. Comparative analysis includes analysis of repeated sequences in the reference human genome, structural variation across the human population and sequence elements conserved across multiple species. The annotation process for functional genomics data involves smoothing the raw signal (step 1), thresholding and segmentation of the smoothed signal (step 2), clustering of discrete segments (step 3), functional annotation of clusters (step 4) and connecting clusters into networks (step 5).

Here, we describe how a similar process can be applied to genome annotation: we can do large-scale similarity comparisons on the genome sequence, taking note of repeated regions at different scales, and then look for function in the genome by mapping the 'read-out' from experiments onto sequence elements. The process for annotating the human genome can be separated into the two broad categories of comparative sequence analysis (comparative analysis) and functional genomics analysis (functional analysis), which correspond to analysing DNA sequences and analysing the output from functional genomics experiments, respectively (FIG. 1). In this Review we will focus mainly on functional analysis and will provide only a brief overview of comparative analysis simply as a framework for showing how it can be integrated with functional analysis.

Comparative analysis

The field of DNA sequence analysis is in the middle of a paradigm shift caused by the exponential reduction in the cost of obtaining genome sequence data. The traditional scope of comparative genomics is the comparison of reference genome sequences from different species; however, the recent explosion in sequencing has made it possible to sequence populations of a species and, in cancer genomics, the genomes of normal and diseased cells within an individual. Similar concepts and sequence analysis tools apply whether one is comparing one human genome sequence to itself, to that of another human or to that of another species. Here, we use 'comparative analysis' to encompass all of these activities, as a more specific term is lacking in the field.

Repeated sequences that can be identified in the reference human genome include segmental duplications (also known as low-copy-number repeats (LCRs)), simple and tandem repeats, transposons and pseudogenes. We emphasize that here the term 'repeated sequence' refers to a wider set of elements than the term 'repeat element', which typically refers to a short, highly repetitive sequence. Structural variants are revealed by comparing genome sequences across the human population, and conserved NCEs, large syntenic blocks and orthologous genes are revealed by comparing the human genome to those of other species (BOX 1; TABLE 1). There are two main methods for discovering repeated sequences: first, scanning for sequence similarity, which involves grouping together sequences that fall above a minimum threshold of sequence conservation over some length scale; and second, model-based discovery, in which curated sets of known elements form the basis of statistical models which are then used to scan the genome for additional elements that fit the model[25]. Note that these two approaches can be used to compare the sequences of different organisms (in the traditional sense of comparative genomics), to compare the sequences of organisms within a population (in the sense of personal genomics) or to compare sequences within an individual (when comparing cancerous to normal cells).

Structural variants
Chromosomal rearrangements (deletions, duplications, novel sequence insertions or inversions) that are inherited and polymorphic across the human population. Structural variants are by definition longer than SNPs and can be hundreds of thousands of base pairs long.

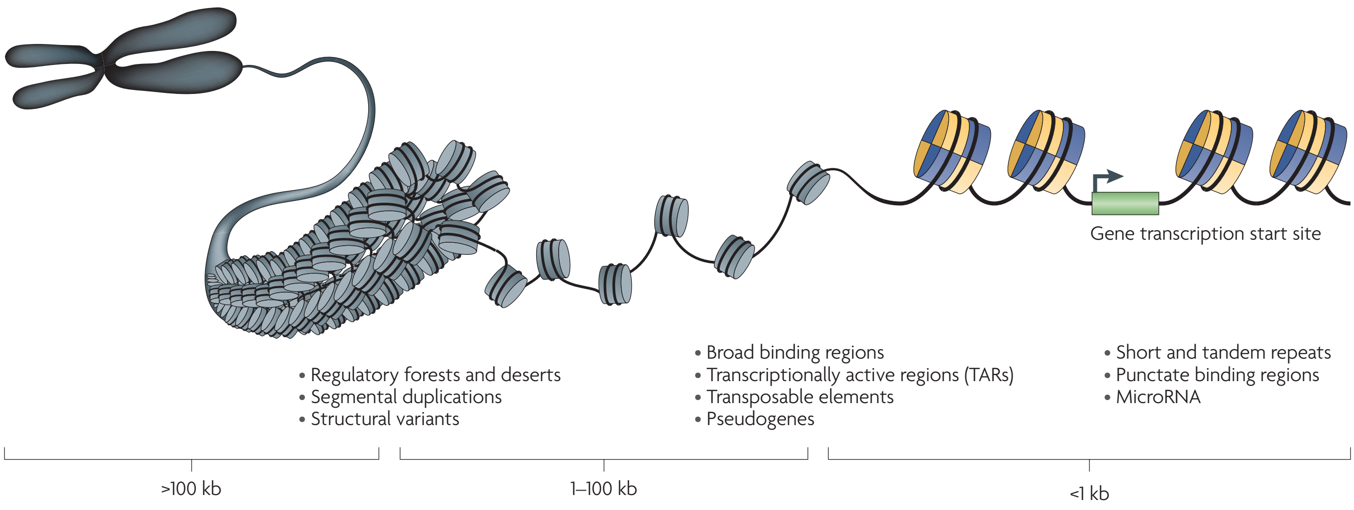
Copy-number variants
Structural variants that arise from deletion or duplication and thus lead to a change in copy number of the underlying region of the genome.

Segmental duplication
The operational definition of a segmental duplication rests on finding two regions in the same genome ranging in length from a thousand to several million nucleotides with at least 90% sequence identity. Segmental duplications are inherited but not necessarily polymorphic across the human population.

Pseudogenes
Copies of protein-coding genes with mutations that disrupt their coding sequence and demolish their original protein-coding function.

Syntenic blocks
Segments that align between genome sequences from two species and that are believed to define an orthologous relationship.

Box 1 | Catalogue of non-coding elements



Non-coding elements are found at all scales within the genome: short repeats, regulatory factor binding regions and small RNAs at small scales; broad histone marks, transcripts, transposable elements and pseudogenes at medium scale; and regulatory forests and deserts, segmental duplications and structural variants at larger scales (right to left in the figure). Insertions and deletions in the human genome range in scale from SNPs to chromosome-scale abnormalities.

Simple and tandem repeats

Simple repeats are duplications 1–5 bp in length that are probably generated by polymerase slippage errors[100]. Tandem repeats are 100–200 bp duplications and are often found at the centromeres and telomeres of chromosomes, where they have a structural role[101]; variation in their number in gene promoters can affect nucleosome positioning and gene expression[102].

Transposable elements

Transposable elements are divided into DNA-based transposons and RNA-based retrotransposons. Some are still active in genomes today, whereas others have become inactive[103]. Long interspersed elements (LINEs) are retrotransposons that themselves encode reverse transcriptase. Active LINEs increase genome size by copying themselves into new locations. Short interspersed repeats (SINEs) like the human *Alu* element are fragments of RNA polymerase III-transcribed genes that rely on LINE elements for propagation. Long terminal repeat (LTR) retrotransposons are flanked on both ends by direct LTRs. They become inactive when homologous recombination between the LTRs deletes the intervening *gag* and *pol* genes[8,9].

Pseudogenes

Several categories of pseudogenes have been annotated, including duplicated pseudogenes, processed pseudogenes and unitary pseudogenes[73,104].

Segmental duplications

About 45% of human segmental duplications occur in tandem runs spaced less than 1 Mb apart on the same chromosome[71].

Structural variants

These can be generated by insertion, deletion, reciprocal translocation or inversion[17,18]. Duplications and deletions cause copy-number variation across the population.

Conserved and ultraconserved non-coding elements

Comparative genomics has found non-coding elements (NCEs) that are conserved to varying degrees across mammalian or vertebrate genomes, which suggests some function conserved by natural selection[94,12,13]. Lengths of conserved NCEs range from one to thousands of base pairs. Lack of function for some ultraconserved elements casts doubt on the assumption that sequence conservation implies function[11,12,78].

Functional non-coding RNAs

Recent years have seen a revolution in our understanding of the role of small regulatory RNAs[10]; new classes continue to be discovered. MicroRNAs (miRNAs) are 22 nucleotide (nt) RNAs that bind predominantly to the 3′ UTRs of mRNA, causing gene silencing[105–107]. Small interfering RNAs (siRNAs) are 21 nt long and also function in the degradation of complementary mRNAs[10]. Piwi-interacting RNAs (piRNAs) are 27–28 nt RNAs that repress transcription of transposons in the germ line of fruitflies[108] and vertebrates[109]. Large intergenic non-coding RNAs (lincRNAs) are spliced like protein-coding genes that function in several central cellular processes[81,82]. Small RNAs, such as small nucleolar RNAs (snoRNAs), generated by RNA polymerases I and III help to synthesize the translational apparatus and make up 90% of the RNA in the cell.

Regulatory elements

The human genome contains 1,700–1,900 transcription factors[110]. Binding sites of some 100 transcription factors have been characterized at genome scale by chromatin immunoprecipitation followed by microarray (ChIP–chip)[20,33,34] or by sequencing (ChIP–seq)[36,37]. Classes of regulatory elements to which transcription factors bind include promoters, enhancers, silencers, insulators and locus-control regions (LCRs)[20,111]. Promoters are regulatory sites that alter the expression of the nearest gene, whereas the other elements act on more distant genomic locations.

DNA-based transposons
Transposable DNA elements that rely on a transposase enzyme to excise themselves from one region of the genome and insert themselves into a different region, without increasing in copy number.

Scanning for sequence similarity. Looking for regions of sequence similarity consists of grouping sequences that share some minimum sequence similarity over a specified minimum length. The main problem with this approach is that it invariably leaves out related regions that have degraded over time, so their similarity is below the threshold. Moreover, the thresholds chosen to group elements together often have no connection to evolutionary history and the underlying mechanisms of formation. For example, the operational definition of segmental duplications (BOX 1; TABLE 1) excludes ancient duplications that were formed by the same mechanisms long ago but that have since degraded below 90% sequence identity.

Table 1 | **Length, number and genome coverage of a representative collection of non-coding features**

Classification	Property	Length (nucleotides)		Number of items	Genome coverage (Mb)	Genome coverage (%)
		Average	Longest			
From comparative analysis						
Short and tandem repeats	Simple repeat	63	2,961	415,917	26.1	0.84
	Satellite	1,444	160,602	8,997	13.0	0.42
	Low complexity	46	2,023	370,102	17.0	0.55
DNA transposons		215	3,625	459,524	98.6	3.17
Retrotransposons	LINEs	426	8,505	1,490,241	634.6	20.4
	Alu SINE element	261	614	1,186,885	309.7	9.97
Pseudogenes	Duplicated	6,607	181,882	2413	15.9	0.51
	Processed	723	15,732	8303	6.0	0.19
Segmental duplications		5,740	630 kb	26,469	151.9	4.89
Structural variants		8,761	3.3 Mb	96,874	848.8	27.3
From functional analysis						
Punctate binding sites	STAT1	446	9,079	~2,300	1.0	0.03
	CTCF	1,181	79,200	~35,000	41.4	1.33
	H3K4me3	1,759	71,025	~62,000	110.2	3.55
Broad binding sites	H3K36me3	4,518	380,076	~130,000	589	19.0
MicroRNA		89	150	718	0.063	0.00
TARs		72	1,854	644,200	46.7	1.50
Regulatory forests		3,890	35,165	68,900	268	8.62
Regulatory deserts		27,107	203,691	72,500	1,970	63.4

Pseudogene counts are taken from build 53 at Pseudogene.org[29]. MicroRNA counts are from miRBase[121]. Counts of structural variants are from the Database of Genomic Variants[122]. Data on transcriptionally active regions (TARs) and regulatory forests and deserts are extrapolated to whole-genome scale from the 1% of the genome covered by the ENCODE pilot project[20]. The extrapolation is biased by the high fraction of genic regions in the ENCODE pilot regions. All other data were collected from the University of California-Santa Cruz (UCSC) Table Browser[45] using the March 2006 build of the human genome (UCSC hg18, NCBI build 36). CTCF, CCCTC-binding factor; H3K4me3, histone 3 lysine 4 trimethylation; H3K36me3, histone 3 lysine 36 trimethylation; LINE, long interspersed element; SINE, short interspersed element; STAT1, signal transducer and activator of transcription 1; TAR, transcriptionally active region.

Model-based discovery of non-coding elements. Some classes of element can be identified by using more sensitive, model-based comparison techniques. In particular, in situations in which more detailed information about the structure or mechanism of formation of a specific element is available, we can use it to discover more diverged class members[25]. We can also search for elements based on their tendency to fold into stable structures[26].

Transposable elements and pseudogenes are examples of non-coding sequences that can be identified by using models based on the descent of these sequences from protein-coding elements. For instance, the same powerful tools used to identify protein-coding genes can be used to identify active transposable elements that still code for (retro)transposase enzymes. Inactive transposable elements can be identified by their similarity to active transposable element profiles and by the stereotypical structure of short repeats at their margins that mark excision scars. Likewise, protein sequence similarity to parent genes is the main feature used to detect pseudogenes[27–29] and is a much more sensitive indicator than raw nucleotide identity.

Where previous work has identified a set of genes that are all regulated by the same regulatory factor, statistical tools, such as Gibbs sampling[30], can reveal subtle motifs that are common to the promoter and enhancer regions of all of the genes to which the regulatory factor binds. Scanning the genome with a model of such a sequence motif can identify a more complete set of binding regions.

This brief overview of comparative analysis is intended simply to provide a context for the following functional analysis section. Readers who wish more detail are referred to several excellent reviews[30–32]. In the remainder of this Review, we focus on functional analysis and its integration with comparative analysis.

Functional analysis

In functional genomics, experimental techniques that characterize the biological role of genetic sequences are expanded to generate data at genome scale in a high-throughput way. For instance, chromatin immunoprecipitation followed by microarray (ChIP–chip)[33–35] or by sequencing (ChIP–seq)[36–38] can be used to identify regulatory-factor-binding regions (RFBRs), and transcription tiling arrays[39,40] and RNA sequencing (RNA–seq)[41–44] can be used to identify transcriptionally active regions (TARs). Here, we give an overview of a standardized

RNA-based retrotransposons
Transposable elements generated when reverse transcriptase enzymes copy RNA elements into DNA and insert the DNA copies back into the genome.

Duplicated pseudogenes
Pseudogenes that result from whole-genome or segmental duplications, in which one copy maintains its ancestral function and the other copy degrades into a pseudogene.

Processed pseudogenes
Pseudogenes that arise when the mRNA of a parent gene is retrotranscribed back into DNA and inserted into the genome.

Unitary pseudogenes
A rare class of pseudogene in which a single-copy parent gene becomes non-functional.

Chromatin immunoprecipitation
(ChIP.) A technique for identifying potential regulatory sequences that are bound by the protein of interest. Soluble DNA–chromatin extracts (complexes of DNA and protein) are isolated by using antibodies that recognize specific DNA-binding proteins. In ChIP–chip, the ChIP step is followed by microarray analysis, whereas in ChIP–seq, it is followed by sequencing.

Tiling arrays
A class of microarray in which probes of a specific length and spacing provide uniform coverage of an entire genome or portion of a genome to a desired resolution.

RNA sequencing
The use of high-throughput sequencing of RNA that has been reverse-transcribed into DNA to characterize the set of RNA transcripts produced by a cell.

Figure 2 | **Signal resolution and signal thresholding. a** | Comparison of signal tracks obtained from chromatin immunoprecipitation followed by sequencing (ChIP–seq) and ChIP followed by microarray (ChIP–chip). The example shown focuses on the binding of the transcription factor signal transducer and activator of transcription 1 (STAT1) to the promoters of genes in the interleukin receptor cluster on chromosome 21. It is clear that following ChIP with short read DNA sequencing (ChIP–seq, top) generates a much cleaner signal than using a microarray (ChIP–chip, bottom). The ChIP–seq track clearly identifies three STAT1 binding sites, whereas the ChIP–chip track requires a more complex thresholding step. There is negative signal (red) in the ChIP–chip track because the microarray signal is a ratio of STAT1 binding compared to a control state. Positive binding signals of the same magnitude should be treated as noise. **b** | Issues with signal thresholding. The genes stromal antigen 2 (*STAG2*) and X-linked inhibitor of apoptosis (*XIAP*) have different levels of exonic, intronic and intergenic transcription signals in a tiling array signal track. If the global threshold to differentiate signal from noise is set high (stringent threshold), exons and introns in highly expressed genes (here *STAG2*) will be correctly segregated, but even exons of weakly expressed genes (here *XIAP*) will not be flagged as expressed. Conversely, if the threshold is set low enough (relaxed threshold) to differentiate exons from introns in weakly expressed genes (*XIAP*), then both introns and exons of highly expressed genes (*STAG2*) will be flagged. These difficulties in thresholding can lead to intronic RNA from precursor mRNA being flagged as expressed transcriptionally active regions (TARs). *IFNAR*, interferon (alpha, beta and omega) receptor; *IL10RB*, interleukin 10 receptor, beta.

Smoothing
The process of filtering noise from a signal by removing fine-scale variation.

Thresholding
The process of discretizing a continuous signal by choosing a signal value above which the signal is considered 'on' or 'active' and below which the signal is considered 'off' or 'inactive'.

Segmenting
The result of thresholding in signal processing — that is, segments are those regions defined as 'on' or 'active' after discretization of the signal.

signal processing approach to analysing such functional genomics data sets. We do not review the long history of annotation of NCEs on an element-by-element basis (BOX 1; TABLE 1).

A useful way to conceptualize the analysis of a generic functional genomics experiment is with a signal-processing paradigm. Each experiment generates a raw signal of some kind across the genome that can be analysed by smoothing it and then thresholding and segmenting it into discrete units of initial annotation. In practical terms, the ubiquity of this paradigm is apparent from the fact that the University of California-Santa Cruz (UCSC) Genome Browser, a major clearing-house for genomic information[45], treats each experiment as a separate 'signal track'. A signal track usually represents a continuous-valued number across the genome, which we can transform into a set of discrete genomic regions, or 'hits', represented in another track. First, we explain a signal-processing pipeline that transforms raw signal tracks into processed annotation tracks. Later, we highlight how integrative analysis of multiple tracks can lead to larger, derived annotations.

Primary data processing: smoothing the raw signal. The raw signal of a functional genomics experiment gives the read-out of transcription, protein binding or some other biological process at discrete points in the genome. Depending on the technology used, signals are mapped to the reference genome with different resolutions. High-throughput sequencing generates alignments at base-pair resolution, whereas tiling arrays provide resolutions from 5 to 50 bp, depending on probe density[39,40] (FIG. 2a). The output is a noisy signal consisting of many piled-up sequence reads or probe values. From this noisy data, the goal is to determine where a transcription factor actually binds[36,37], where a particular DNA[46] or histone[47–49] modification is being made, or what sequence is being transcribed[39–42].

Numerous technical issues that influence signal quality need to be addressed. In particular, because arrays rely on hybridization to measure the amount of target DNA present, the signal obtained for each oligonucleotide probe is modulated by its sequence composition. Probes with greater GC content, for example, show higher signal[50]. Another issue is cross-hybridization, in which regions of the genome with similar sequences bind to multiple probes on the array. Cross-hybridization often gives rise to spikes in the signal, causing problems for measuring the expression of multi-gene families and pseudogenes with tiling arrays. Sequencing technologies do not suffer from cross-hybridization; however, analogous problems occur because short sequence reads can misalign to an incorrect location in the genome owing to sequencing and mapping errors[51–54]. In general, correct read mapping is one of the main technical challenges in next-generation sequencing (BOX 2).

Thresholding and segmenting to generate small initial annotations. After smoothing, it is necessary to set a threshold to differentiate regions with and without signal. Thresholding issues have been most thoroughly explored for ChIP–seq experiments, which we focus on here. We expect that approaches for thresholding RNA–seq signals will evolve along similar lines.

To correctly construct local thresholds, it is important to model or simulate an appropriate null process for the background[55,56]. Because the background signal can be noisy (FIG. 2b), naive methods of thresholding using the assumption of a uniform background are not successful.

In ChIP–seq experiments, the signal from ‘input DNA’ is often used as the background. This signal is generated by sequencing genomic DNA without any enrichment step. The most commonly accepted explanation for the non-uniformity of this control signal is that it reflects the chromatin state of the genome. Regions of open chromatin are more likely to shear and generate DNA fragments of an appropriate size to pass a sizing filter and be captured by sequencing[57]. Using the input DNA signal as background also accounts for the differential ‘mappability’ of regions of the genome — that is, the fact that some regions, most obviously repeats, are underrepresented in the output of the experiment because they are less likely to produce reads that can be mapped uniquely back to the reference genome.

The initial output from thresholding and segmenting an experimental signal is a number of small annotation blocks that are represented as a discrete ‘feature’ track[45]. The next step is to assign biological meaning to the blocks. The experimental read-out is interpreted differently depending on whether the experiment involves transcription or immunoprecipitation.

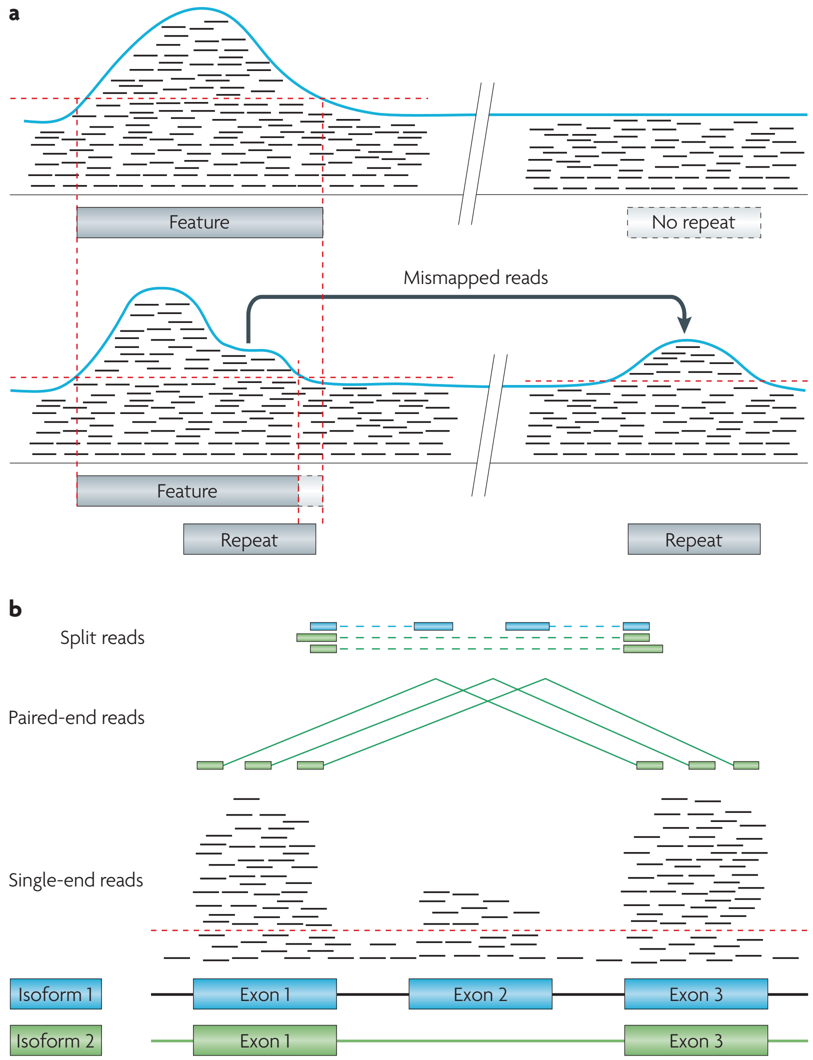
Interpreting the initial annotations: transcriptionally active regions. The result of thresholding RNA–seq or tiling array signals[58,59] is a set of TARs (also known as transcription fragments (‘transfrags’)). Although most TARs stem from protein-coding genes, they can also mark non-coding RNAs. An unexpected result from the ENCODE pilot project was the discovery of pervasive transcription — that is, large numbers of novel TARs in unannotated portions of the genome[20,60]. There is much debate over whether these and other unannotated transcripts are functional or simply the result of cross-hybridization or transcriptional noise[61–63]. Although the fraction of transcribed RNA sequences that map to intergenic and intronic regions is fairly low (~5–10%), that set of TARs covers a relatively large fraction of nucleotides in the genome. This finding is consistent with the fact that annotated genic regions are transcribed at higher levels. Moreover, even though a large fraction of the human genome is transcribed as primary transcripts, which include introns, it remains a challenge to distinguish novel processed RNA products from remnants of primary transcription that can be associated with known genes[63].

Interpreting the initial annotations: regulatory factor binding. Segmentation of ChIP–chip or ChIP–seq signals generates RFBRs[64]. This awkward abbreviation was chosen by the ENCODE consortium to refer to both transcription factor binding and histone modification experiments, as both are important for genome annotation. (Other abbreviations considered include CHIRP (chip hit of regulatory potential) and EIGR (experimentally identified genomic region).) Here, we use the simpler term ‘binding sites’ for RFBRs and highlight several issues with transforming a set of raw binding sites into more developed annotations.

First, binding sites can be divided into two major classes: punctate and broad. For example, some histone modifications cluster in sharp peaks around transcript promoter regions (punctate binding sites), whereas others mark the entire transcribed region with a broad peak (broad binding sites)[49]. Although punctate sites have, for a long time, been identified by scoring algorithms, methods for identifying signals across broad regions of the genome have been less thoroughly developed[65,66].

Second, binding sites differ in the degree to which they have a clear sequence motif connecting them to their associated transcription factor. A transcription factor with a weak motif may be present at very high concentration in a given tissue, binding more promiscuously and activating more genes than in another tissue in which it is present at lower concentration[67]. Even for transcription factors with a strong motif, chromatin accessibility often modulates binding in a cell-type-specific manner[47,48]. A strong binding event may require both open chromatin and a matching motif. A complete picture of transcription factor binding thus needs to incorporate both sequence information about transcription-factor-binding-site motifs and functional genomic information about chromatin state and transcription factor expression level.

Box 2 | **Assembling and mapping repeat regions of the genome**



When the human reference genome was finished in 2004, 340 gaps containing an estimated 200 Mb of heterochromatin and 25 Mb of euchromatin were still not sequenced because of the highly repetitive nature and difficulty of assembling those regions[112]. Current efforts focus not on completing the human reference genome but on supplementing it with data from individuals from diverse populations. The task of generating a human reference genome has transformed into one of thoroughly cataloguing structural variants in the human population.

For reasons of cost, human population genomics efforts, such as the 1000 Genomes Project, rely on short-read DNA sequencing[113]. Detecting repeat regions and structural variants from short reads is a key technical challenge because structural variants are enriched with repeated sequences that map to multiple locations in the genome when probed by short reads. In simulations, multiple algorithms designed to detect structural variants from short reads do not find up to half of the structural variants known from traditional capillary sequencing of Craig Venter's genome[114–117]. A useful resource generated by the Human Genome Structural Variation Project to deal with this problem is a library, generated from multiple individuals, of 40-kb long fosmid clones, the ends of which have been sequenced by paired-end Sanger sequencing[18,118]. A subset of so-called discordant clones with spans that, when mapped onto the reference genome, are substantially longer or shorter than expected (owing to deletions and insertions, respectively) was probed by tiling arrays[18]. Several clones from this subset were also probed by Sanger sequencing of the whole clone[116,118]. It might be possible to do both the paired-end sequencing and clone assembly using short-read sequencing, resulting in a set of 40 kb contigs that could be easily mapped to the reference genome.

After the set of structural variants across the human population has been characterized in sufficient depth, it may be possible to identify and genotype structural variants in an individual genome using short-read sequences and a library of structural variant breakpoints from the whole population[119]. It was recently estimated that a set of human reference genomes covering repeat regions and structural variants for all major human populations might include 19–40 Mb of novel DNA sequence that is not present in the current reference genome[120].

The figure shows the mapping of short reads onto the genome. Mappability is a problem for short read data sets (part **a**). Short reads (black horizontal lines) generated by repetitive elements of the genome can map to multiple locations, generating ambiguity in read counts for highly repetitive, poorly mappable regions. Features that contain repetitive sequence can suffer from mismapping of short reads to other genomic locations with high sequence similarity. Part **b** shows the way in which connectivity maps can be generated between widely spaced regions of the genome. In genes and long non-coding RNAs, split-read and paired-end read methods can identify exon junctions better than noisy single-end read data. These methods also enable the identification and quantification of alternative isoforms, in addition to being useful for identifying structural variants (not shown).

Heterochromatin
Highly compact and therefore inactive regions of the genome. Largely composed of repetitive DNA, heterochromatin forms dark bands after Giemsa staining.

Euchromatin
The lightly staining regions of the genome that are generally decondensed during interphase and contain transcriptionally active regions.

Fosmid
A low-copy vector for the construction of stable genomic libraries that uses the *Escherichia coli* F-factor origin of replication. Each fosmid clone can store ~40 kb of library DNA. Cloned sequences are more stable in fosmids than in high-copy vectors.

One difficulty connected with determining transcription factor binding motifs is the loss of information on binding specificity owing to crosslinking. Transcription initiation complexes often consist of a DNA element bound by multiple interacting transcription factors, some bound to distal enhancer regions that are adjacent in three-dimensional (3D) space because of chromosomal looping. Thus, immunoprecipitation of one transcription factor in such a complex may elute DNA to which it binds only indirectly. The resulting set of target sequences will identify poorly the sequence motif to which the transcription factor actually binds[68].

Third, determining the relationship between binding sites and their target genes is crucial to gaining a global picture of transcriptional regulation, including epigenetic mechanisms. Moreover, this information is the starting point in building regulatory networks that connect transcription factors with their targets. In compact genomes, such as those of yeast and *C. elegans*, associating binding regions with downstream targets is fairly straightforward. However, in the vast expanse of the human genome, this determination is less straightforward. Sites that are thousands of bases apart are often brought into proximity by complex chromatin structures, including looping.

Integrating information

The types of data presented above can be displayed as a single track in a genome browser. This track can show either a continuous signal across the genome or a set of discrete 'hit' regions, such as binding sites. The next step is to group the information from a single track or from multiple tracks into larger annotation structures, such as entire transcripts, that have more biological meaning. Eventually, multiple classes of functional elements that are not proximally located on the genome can be wired together into networks.

Grouping small annotation units into larger structures with a genomic matrix. Track integration begins by generating a 'genomic matrix' in which each row corresponds to a different experiment and each column to a different genomic region (FIG. 3). Then each matrix cell represents the aggregated read-out of a particular experiment within a specific region — for example, the average transcriptional signal within a specific 1 kb region of the genome in the HeLa cell line or the number of nucleotides in that region bound by a specific transcription factor. The genome can be decomposed naturally into regions of different scales (BOX 1; TABLE 1); correspondingly, different matrices can bin genomic regions at different resolutions.

Simple statistical operations on genomic matrices can then provide useful information. In particular, for a set of tracks of experimental features binned at a fine resolution (for example, factor binding sites collected in small 1 kb bins), one can find larger blocks (for example, 150 kb) that are statistically enriched or depleted in these features compared with a randomized null distribution. Enriched and depleted regions have been termed 'regulatory forests' and 'regulatory deserts'[69] (FIG. 3Aa).

Correlated genomic regions can be identified by clustering columns of the matrix[70]. For example, to identify groups of co-regulated TARs, one can build a matrix in which the rows correspond to different cell lines or tissues with transcriptional information and the columns correspond to transcribed regions (including exons) that have been identified in any of the experiments. We can compute correlations of the column vectors of expression signals between novel TARs and nearby known exons. Novel TARs co-expressed with exons of neighbouring genes are likely to be part of the same larger transcriptional unit[70]. In addition, novel TARs distant from any known gene can be clustered into groups with strongly correlated expression signals, which can help in piecing together larger non-coding transcript structures. (This operation can be compared with the connectivity provided by paired-end reads, which are described below.)

Clustering at a higher level naturally gives rise to networks of transcripts that are co-expressed across cell lines or other conditions. The same kind of column clustering applied to transcription factor binding sites, for example, would form a network of co-regulated target genes (FIG. 3Ab).

Co-clustering approaches: biplot. The next step is to examine simultaneous clustering of columns and rows. For example, we can recognize pairs of factors that often bind together by the high correlation of their row vectors in the genomic matrix. Computing the correlation of each factor against all others gives rise to another matrix, called the correlation matrix (FIG. 3Ac). Likewise, we can also cluster regions of the genome (matrix columns) together based on which factors bind them, which generates a second correlation matrix for regions (FIG. 3Ab). Principal components analyses of these two correlation matrices give rise to 'eigen-factors' (which represent the typical behaviour of factors across the genome) and 'eigen-regions' (which represent the typical modes of binding of many factors across the genome).

Given a data matrix consisting of the number of times each factor binds each region, we can cluster regions with regions and factors with factors, and cross-correlate regions with factors. All three of these types of linkage can be visualized in a biplot[69], which shows region and factor clustering simultaneously (FIG. 3Ad). Effectively, the biplot performs principal components analysis on each of the correlation matrices and shows the natural interrelationship of eigen-factors and eigen-regions.

Aggregation and saturation plots. Another type of analysis that can be done with the genomic matrix is the saturation plot. Here, one looks at the cumulative fraction of the genome 'covered' by adding more rows to the matrix to see how many different assays one needs to achieve 'saturation' of a particular type of element. This type of plot has been used extensively in the ENCODE and modENCODE projects to measure overall progress in annotating the genome. One complication with saturation plots is that the slope of the increase in cumulative fraction depends on the order in which the assays are chosen. To get around this problem, one can shuffle the assays and show a box plot resulting from all different possible orderings (FIG. 3B).

Specificity
A measure of the proportion of true negatives correctly identified as such (for example, the percentage of healthy people who are identified as not having a disease).

Regulatory forests
Regions of the genome that are enriched with binding sites for regulatory factors, such as transcription factors.

Principal components analysis
A statistical method used to simplify data sets by transforming a series of correlated variables into a smaller number of uncorrelated factors.

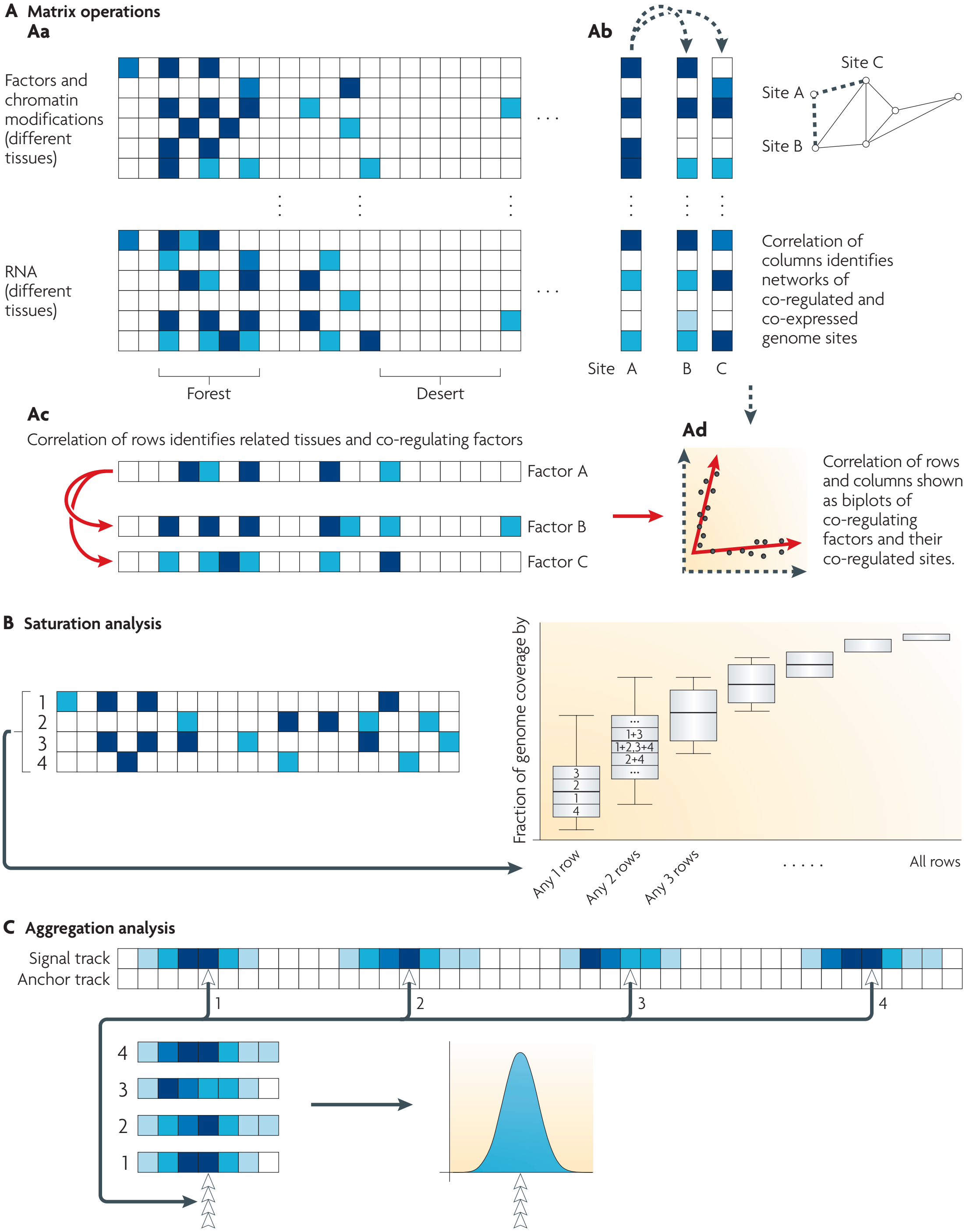


Figure 3 | **Matrix showing how to correlate genomic elements. A** | Simple matrix operations. Each row in the matrix corresponds to a different experiment and each column to a different genomic region (**Aa**). The numerical value of each matrix element corresponds to an aggregated read-out of that experiment in that specific region of the genome. Simple statistical operations on the matrix can provide useful information. For example, correlations between columns (**Ab**) identify networks of co-regulated and co-expressed genome sites, whereas correlations between rows (**Ac**) identify related tissues and co-regulating factors. Simultaneous correlation of rows and columns (**Ad**) can associate co-regulating factors with the sites they regulate. Grouping columns into regions enriched or depleted for regulatory sites compared with the genome average identifies regulatory forests and deserts. **B** | This schematic saturation plot shows how genome coverage increases as related signal tracks are joined together. Here, signal track 3 covers a larger fraction of the genome than any other single track; signal tracks 2 and 3 together cover more of the genome than any other pair of tracks, and so on. Saturation plots present genomic summary statistics in a useful visual framework. **C** | This schematic aggregation plot shows how a class of genomic features from one annotation track can be used as anchor points to sum up the values in a related set of signal tracks. This plot could represent, for example, the average profile of short RNA sequencing reads around all transcription start sites in the genome.

In addition to measuring saturation across tracks, one can analyse the statistical distribution within a single track using an aggregation plot. Here, one sums the signal within a set distance around all instances of a set of genomic anchor points — for example, around transcription start sites or within exons (FIG. 3C). In a sense, each aggregation plot builds a special coordinate system for the matrix in which matrix elements (or bins) are placed at predefined distances from each of the genomic anchor points, which can be expressed as another track.

Analysis of sequence features in a similar framework. The type of integrated analysis done for different classes of functional genomics tracks can also be done for tracks defined by sequence analysis. For example, we might expect some correlation between the occurrence of segmental duplications and short repeat elements, as one of the formation mechanisms of segmental duplications is non-allelic homologous recombination (NAHR)[71], and the presence of short repeats increases the likelihood of non-allelic crossing-over. This correlation uses the same genomic matrix approach described above, except bins represent the number of segmental duplications or short repeats (for example, *Alu*s) in a genomic interval. In fact, one study that related segmental duplications to short repeats showed that segmental duplications tend to be associated with *Alu*s and that the change in this association over time highlights the effect of the *Alu* burst ~40 million years ago[72].

Integrating comparative and functional tracks

In the previous section, we looked for correlations between regions of the genome that shared either sequence-based or experimental features. Here, we combine information from both comparative and functional analysis. We can do this in two ways. First, we can measure the overlap between the two sets of features in terms of the number of base pairs. Using the genomic matrix framework, we can compute a correlation between the rows of the matrix that represent sequence features and the rows of the matrix that represent functional features. Second, we can calculate a 'sequence metric', such as the degree of conservation or variability, for each functionally annotated feature or a 'functional metric', such as the amount of transcription, for each sequence feature.

When calculating such metrics, it is important to assess their values relative to an appropriate genomic null. For example, is a functional element more or less conserved than one would expect it to be by chance? One can determine this expectation trivially by randomly shuffling elements in the genome. However, there are a number of better ways to construct an appropriate null, such as by using the genome structure correction (GSC) statistic[20].

Below, we highlight these approaches to interrelating comparative and functional analysis using a number of representative case studies.

Detecting transcribed pseudogenes. An example that combines comparative and functional evidence is the annotation of transcribed pseudogenes. Pseudogenes identified by sequence similarity to parent genes — fundamentally a result of comparative analysis — can be examined for transcriptional activity by comparison with functional tracks derived from RNA–seq or tiling-array data. In fact, evidence from the ENCODE pilot project suggests that at least 20% of human pseudogenes are transcribed[73]. It has been suggested that some transcribed pseudogenes have been recruited into the RNA-interference pathway to control transcription of their parent genes. In these cases, the antisense transcript from the pseudogene binds to the mRNA of its parent, generating a natural endogenous small interfering RNA[74,75]. These observations suggest that pseudogenes can play a significant part in gene regulation[76]. However, in the ENCODE pilot no obvious sequence signature for transcribed pseudogenes could be found — that is, they were conserved no more and have no fewer SNPs than other pseudogenes[73].

Sequence conservation versus function. This ambiguous finding about transcribed pseudogenes is an example of the broader result from the pilot ENCODE project that conserved elements identified by comparative analysis are not always functional, and vice versa[20,77,78]. Before the project, it was expected that, to some degree, all conserved blocks would have some function mapped to them. Somewhat surprisingly, many blocks were found to have no experimental evidence of function and, conversely, many experimentally identified functional elements were not conserved. We highlight the case of ultraconserved elements. Some of these elements are tissue-specific enhancers[11,13], whereas others have a role in mRNA degradation[79]. However, deletion of several ultraconserved elements in mice was not lethal and caused no problems in growth, longevity, fertility or metabolism[77]. If some function is not found for conserved non-coding regions even after an exhaustive array of functional assays, evolutionary models may need to be revised[78,80].

Annotating lincRNAs. A final example of integrating comparative and functional analysis comes from the study of large intergenic non-coding RNAs (lincRNAs; also known as large intervening non-coding RNAs). 'K4-K36 domains' are histone signatures that mark actively transcribed elements with a punctate histone 3 lysine 4 trimethylation (H3K4me3) mark at the transcription start site and a broad H3K36me3 mark across the transcribed region[81]. A large set of K4-K36 domains identified from functional genomics experiments in mice was screened against known protein-coding genes and regulatory RNAs to find domains without any known annotation[81]. A custom tiling array designed to map a subset of those unannotated regions showed transcription in most of them. Sequence analysis showed that almost none of the transcribed elements was protein-coding, so they represented a set of candidate lincRNAs. When clustered with protein-coding genes of known function by their shared expression level across several tissues, groups of lincRNAs involved in the DNA damage response, immune signalling and maintenance of stem cell pluripotency were identified. Recent work has extended this analysis into humans[82].

Non-allelic homologous recombination
Recombination between segmental duplications that leads to local duplication, deletion or inversion of genome sequence.

Ultraconserved elements
Operationally defined as non-coding elements that are hundreds of base pairs long and 100% identical across human, mouse and rat genomes.

Table 2 | **Percentage of non-coding DNA in selected sequenced genomes**

Species name		Genome size (Mb)	Fraction of genome (%)				Source of gene annotations
Common	Scientific		Genic	Exonic	Non-coding		
					Intronic	Intergenic	
Yeast	*Saccharomyces cerevisiae*	12.2	73.5	72.9	0.6	26.6	*Saccharomyces* Genome Database (June 2008 build)
Nematode worm	*Caenorhabditis elegans*	100.3	59.2	28.1	31.2	40.8	WormBase (WS190)
Fruitfly	*Drosophila melanogaster*	168.7	48.2	18.3	30.0	51.8	FlyBase and Berkeley Drosophila Genome Project (BDGP; release no. 5)
Human	*Homo sapiens*	3,107	45.1	2.8	42.3	54.9	UCSC Genome Browser Known Genes table (hg18)

The genic fraction consists of both exonic and intronic sequence. The exonic fraction consists of both coding sequence (CDS) and 5′ and 3′ UTRs. Strictly speaking, UTRs are non-coding, so the exon fraction is a slight overestimate of the fraction of coding sequence in the genome.

Discussion and future directions

We have provided an overview of the annotation process for non-coding regions of the genome. For a long time it has been a mystery why more than 98% of the genomic text seems to have no meaning, with less than 2% consisting of protein-coding exons. The realization that much of the non-coding DNA in the genome is transcribed at low levels into RNA has compounded the mystery[20]. In the future the annotation effort will continue to evolve, driven by rapid improvement in sequencing technologies. Although second-generation platforms sacrifice read length to reduce cost and increase coverage, some third-generation platforms promise to generate very long reads that would span repetitive regions and make assembly much easier[83,84]. We highlight here two directions of future work. One area is a chronic problem that requires attention, and the other is a direction in which the field of functional genomics is already moving.

Validation. The chronic problem area is validation. Validation of predictions from genome-scale experiments using more established molecular biology techniques is crucial. The goal of ENCODE is to make predictions with high accuracy (for example, 5% error rate and 95% sensitivity). A benefit of ENCODE is that researchers can compare different scoring algorithms applied to the same large data sets. However, if each algorithm makes 10,000 binding site predictions for an experiment, perhaps only 60 of them might be targeted for validation by high-quality, low-throughput methods. That number is simply not high enough to readily calibrate the error rates, so it is important that regions for validation are selected systematically to maximize statistical power[85]. On the experimental side, new medium-throughput techniques are under development to increase the number of predictions that can be validated (for example, the NanoString nCounter[86]).

Annotation of connectivity between elements. Most genome-scale data sets discussed in this Review can be displayed in a browser as a single one-dimensional track, whereas new classes of functional genomic data cannot be as easily represented. The key difference is that techniques such as paired-end sequencing[18] and chromosome conformation capture[87] generate connectivity maps that link widely spaced regions of the genome. Additional research is needed to find the most intuitive ways of analysing and visualizing this type of data[88].

In particular, paired-end tag sequencing is beginning to replace conventional single-end sequencing owing to the additional information provided. For RNA–seq, paired-end reads add information about the connectivity of spliced transcripts, including *trans*-splicing events[43,44]. For mapping structural variation, paired reads enable the detection of inversions and translocations in addition to copy-number variation[18] (BOX 2).

Chromosome conformation capture provides information on interactions between DNA elements that are adjacent in the 3D space of the nucleus but located on different chromosomes or widely spaced on the same chromosome[87]. This method involves crosslinking of chromatin, then shearing and ligation to create a library of fused DNA fragments from two distant genomic locations. High-throughput methods, including chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[89], carbon-copy chromosome conformation capture[90] and Hi-C[91], use tiling arrays or deep sequencing to map these fusion products onto the genome. These techniques enable the systematic identification of distant targets of regulatory elements, such as enhancers, and the mapping of the 3D structure of chromatin in the nucleus[91,92].

A paradox of the genomic era has been that the number of protein-coding genes is no higher in humans than in apparently simpler organisms[93]. Human complexity may stem more from differences in regulation than from differences in protein-coding sequences[94]. The ENCODE pilot project showed that non-coding DNA tends to be functionalized mainly in new cell types: transcribed regions conserved across many cell lines were almost exclusively exonic, whereas intronic and intergenic TARs were mainly restricted to a single cell type[20]. So part of the increase in organismal complexity may be caused by the proliferation of cell types. Smaller organisms with fewer cell types have comparatively less non-coding DNA (TABLE 2), although some deviate substantially from this trend[95]. Yeast has three cell types[96]; the nematode *C. elegans* has nearly 1,000 cells of about 20 cell types[97], and humans have perhaps thousands of cell types, not all of them enumerated[98,99].

Sensitivity
A measure of the proportion of true positives that are correctly identified as such (for example, the percentage of sick people who are identified as having a disease).

Paired-end sequencing
Determination of the sequence at both ends of a fragment of known size.

Chromosome conformation capture
A technique used to study the long-distance interactions between genomic regions, which in turn can be used to study the three-dimensional architecture of chromosomes within a cell nucleus.

Future work in understanding the evolutionary trajectory of the human genome should focus on providing functional genomic data in a comprehensive array of human cell and tissue types, a goal that is being pursued by US National Institutes of Health (NIH) projects such as the Genotype-Tissue Expression (GTEx) project and the Transcriptional Atlas of Human Brain Development. Methods currently being developed to annotate NCEs will be critical for the success of such projects.

1. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. *Science* **161**, 529–540 (1968).
2. Ohno, S. So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.* **23**, 366–370 (1972).
3. Lewin, R. Proposal to sequence the human genome stirs debate. *Science* **232**, 1598–1600 (1986).
4. Robertson, M. The proper study of mankind. *Nature* **322**, 11 (1986).
5. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* **106**, 19096–19101 (2009).
6. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotech.* **27**, 182–189 (2009).
7. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
8. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
9. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
10. Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nature Rev. Genet.* **10**, 94–108 (2009).
11. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
12. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
13. Pennacchio, L. A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
14. Kleinjan, D. A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
15. Yeager, M. *et al.* Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum. Genet.* **124**, 161–170 (2008).
16. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
17. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
A prescient exposition of the important link between disease and structural variation in the human genome.
18. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
The first high-resolution sequence map of human structural variation.
19. Lupski, J. R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* **1**, e49 (2005).
20. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
A comprehensive overview of what was learned during the ENCODE pilot project.
21. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
22. Searls, D. B. The language of genes. *Nature* **420**, 211–217 (2002).
23. Whitfield, J. Across the curious parallel of language and species evolution. *PLoS Biol.* **6**, e186 (2008).
24. Pagel, M. Human language as a culturally transmitted replicator. *Nature Rev. Genet.* **10**, 405–415 (2009).
25. Saha, S., Bridges, S., Magbanua, Z. V. & Peterson, D. G. Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.* **36**, 2284–2294 (2008).
26. Washietl, S. *et al.* Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* **17**, 852–864 (2007).
27. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7**, S4 (2006).
28. Zhang, Z. L. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–1439 (2006).
29. Karro, J. E. *et al.* Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.* **35**, D55–D60 (2007).
30. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ. Press, 1998).
31. Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5**, 15–56 (2004).
32. Margulies, E. H. & Birney, E. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nature Rev. Genet.* **9**, 303–313 (2008).
33. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
34. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
35. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protoc.* **1**, 729–748 (2006).
36. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
37. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
38. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
39. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
40. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
41. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA–seq. *Nature Methods* **5**, 621–628 (2008).
42. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
43. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
44. Wang, Z., Gerstein, M. & Snyder, M. RNA–seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
45. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
46. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
47. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
48. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
49. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
50. Royce, T. E., Rozowsky, J. S. & Gerstein, M. B. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics* **23**, 988–997 (2007).
51. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
52. Li, R. Q., Li, Y. R., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
53. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
55. Zhang, Z. D., Rozowsky, J., Snyder, M., Chang, J. & Gerstein, M. Modeling ChIP sequencing *in silico* with applications. *PLoS Comput. Biol.* **4**, e1000158 (2008).
56. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP–seq experiments relative to controls. *Nature Biotech.* **27**, 66–75 (2009).
57. Auerbach, R. K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl Acad. Sci. USA* **106**, 14926–14931 (2009).
58. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
59. Rinn, J. L. *et al.* The transcriptional activity of human Chromosome 22. *Genes Dev.* **17**, 529–540 (2003).
60. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
61. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**, 556–565 (2007).
62. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14**, 103–105 (2007).
63. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most dark matter transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).
A recent reappraisal, based on RNA–seq and tiling-array data, of the degree of pervasive transcription in the human genome.
64. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature Rev. Genet.* **10**, 605–616 (2009).
65. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211 (1998).
66. Gokcumen, O. & Lee, C. Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. *Methods* **49**, 18–25 (2009).
67. Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M. & Levine, M. Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* **111**, 687–701 (2002).
An elegant study of the effect of transcription factor concentration on the arrangement of *cis*-regulatory elements at target genes.
68. Tantin, D., Gemberling, M., Callister, C. & Fairbrother, W. High-throughput biochemical analysis of *in vivo* location data reveals novel distinct classes of POU5F1(Oct4)/DNA complexes. *Genome Res.* **18**, 631–639 (2008).
69. Zhang, Z. D. D. *et al.* Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res.* **17**, 787–797 (2007).
70. Rozowsky, J. S. *et al.* The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. *Genome Res.* **17**, 732–745 (2007).
71. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Rev. Genet.* **7**, 552–564 (2006).
72. Kim, P. M. *et al.* Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* **18**, 1865–1874 (2008).
73. Zheng, D. *et al.* Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res.* **17**, 839–851 (2007).

74. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).
75. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539–543 (2008).
76. Sasidharan, R. & Gerstein, M. Protein fossils live on as RNA. *Nature* **453**, 729–731 (2008).
77. Ahituv, N. *et al.* Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**, e234 (2007).
78. Monroe, D. Genomic clues to DNA treasure sometimes lead nowhere. *Science* **325**, 142–143 (2009).
79. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C. & Brenner, S. E. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446**, 926–929 (2007).
80. Baer, C. F., Miyamoto, M. M. & Denver, D. R. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Rev. Genet.* **8**, 619–631 (2007).
81. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009). **A good example of the benefits of integrating comparative and functional analysis, which in this case led to the discovery of a new class of functional NCEs.**
82. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106**, 11667–11672 (2009).
83. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnol.* **4**, 265–270 (2009).
84. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
85. Du, J. *et al.* A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and ChIP–chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics* **22**, 3016–3024 (2006).
86. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotech.* **26**, 317–325 (2008).
87. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
88. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
89. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
90. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
91. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
92. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010). **References 91 and 92 are two examples of the power of using long-distance connectivity data in the genome to map genome structure.**
93. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA* **104**, 19428–19433 (2007).
94. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
95. Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nature Rev. Genet.* **6**, 699–708 (2005).
96. Galgoczy, D. J. *et al.* Genomic dissection of the cell-type-specification circuit in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **101**, 18069–18074 (2004).
97. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic-cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
98. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81**, 425–455 (2006).
99. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nature Rev. Genet.* **9**, 868–882 (2008).
100. Schlotterer, C. & Tautz, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20**, 211–215 (1992).
101. Amor, D. J. & Choo, K. H. A. Neocentromeres: role in human disease, evolution, and centromere study. *Am. J. Hum. Genet.* **71**, 695–714 (2002).
102. Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K. J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213–1216 (2009).
103. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends Genet.* **23**, 183–191 (2007).
104. Zhang, Z., Frankish, A., Hunt, T., Harrow, J. & Gerstein, M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.* **11**, R26 (2010).
105. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858 (2001).
106. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
107. Lee, R. C. & Ambros, V. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294**, 862–864 (2001).
108. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).
109. Carmell, M. A. *et al.* MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell* **12**, 503–514 (2007).
110. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10**, 252–263 (2009). **A useful synthesis of the current state of knowledge about human transcription factors.**
111. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59 (2006).
112. Bovee, D. *et al.* Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nature Genet.* **40**, 96–101 (2008).
113. Kaiser, J. A plan to capture human diversity in 1000 genomes. *Science* **319**, 395–395 (2008).
114. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, 2113–2144 (2007).
115. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**, 677–681 (2009).
116. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19**, 1270–1278 (2009).
117. Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods* **6**, 473–474 (2009).
118. Kidd, J. M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods* **7**, 365–371 (2010). **The authors report the characterization of new insertion sequences relative to the human reference genome; this study is a useful addition to the field as it moves towards a series of reference genomes for sub-populations.**
119. Lam, H. Y. K. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotech.* **28**, 47–55 (2010).
120. Li, R. Q. *et al.* Building the sequence map of the human pan-genome. *Nature Biotech.* **28**, 57–63 (2010).
121. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008).
122. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).

Acknowledgements

The authors thank members of the Gerstein laboratory for helpful discussions and careful reading of the manuscript. We acknowledge support from the US NIH and from the Albert L. Williams Professorship funds.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Mark B. Gerstein's homepage: http://genometech.gersteinlab.org
1000 Genomes Project: http://www.1000genomes.org
Berkeley Drosophila Genome Project: http://www.fruitfly.org
Database of Genomic Variants: http://projects.tcag.ca/variation
FlyBase: http://flybase.org
GTEx project: http://nihroadmap.nih.gov/GTEx
The ENCODE Project: http://www.genome.gov/10005107
Human Genome Structural Variation Project: http://humanparalogy.gs.washington.edu/structuralvariation
miRBase: http://www.mirbase.org
The modENCODE Project: http://www.modencode.org
Pseudogene.org: http://www.pseudogene.org
***Saccharomyces* Genome Database:** http://www.yeastgenome.org
UCSC Genome Browser: http://genome.ucsc.edu
WormBase: http://www.wormbase.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF