Genome-wide transcription and the implications for genomic organization

Philipp Kapranov, Aarron T. Willingham and Thomas R. Gingeras

Abstract | Recent evidence of genome-wide transcription in several species indicates that the amount of transcription that occurs cannot be entirely accounted for by current sets of genome-wide annotations. Evidence indicates that most of both strands of the human genome might be transcribed, implying extensive overlap of transcriptional units and regulatory elements. These observations suggest that genomic architecture is not colinear, but is instead interleaved and modular, and that the same genomic sequences are multifunctional: that is, used for multiple independently regulated transcripts and as regulatory regions. What are the implications and consequences of such an interleaved genomic architecture in terms of increased information content, transcriptional complexity, evolution and disease states?

Tiling array
A microarray design in which the probes are selected to interrogate a genome with a consistent, pre-determined spacing between each probe.

Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California 95051, USA. Correspondence to T.R.G. e-mail: tom_gingeras@affymetrix.com
doi:10.1038/nrg2083
Published online 8 May 2007

The description of the lac operon in 1961 by Jacob and Monod[1] established a conceptual model of gene organization by which a DNA sequence is neatly split into separate regulatory and protein-coding portions; with the protein-coding portion of a gene preceded by a defined region of DNA that regulates its transcriptional initiation and followed by a functional stretch of DNA that controls its termination. This simple but elegant model has been supported by a wealth of biochemical and genetic data and has consequently become engraved in all thinking regarding genomic organization for nearly every species. A simplistic extension of this model is that a region in a genome usually has just one function; so, a genome consists of a linear arrangement of different functional elements that are interspersed with non-functional elements. For example, a region of DNA can be either a promoter or an exon, but usually not both.

The advent of genome-wide techniques for studying transcription has enabled transcriptome studies on an unprecedented scale (BOX 1). What emerges is that a genomic region can be used for different purposes and that different functional elements can co-locate in the same region in a genome. This observation prompts us to re-evaluate the current dogma, which can be referred to as the 'colinear' model, and indicates an alternative model for genomic organization. This 'interleaved' model reflects the observation that multiple functional elements can overlap in the same genomic space. Here we discuss recent empirical data supporting this and consider the implications, advantages and challenges of this new model of genomic architecture.

Emerging genomic architecture

In-depth analyses of the transcriptional outputs of the human, mouse, fly and other genomes from a range of experimental approaches (TABLE 1; BOX 1) suggest that the information content of a genome is complex, and that this complexity manifests itself at two levels.

The fraction of a genome that is used as an information carrier is much higher than previously expected, and much of the unannotated transcription, the so-called 'transcriptional dark matter'[2], remains to be characterized. Unbiased transcriptome profiling using tiling arrays for ten human chromosomes revealed that 56% of the transcribed base pairs in cytosolic polyadenylated RNA (the cytosol contains the most mature, processed RNAs) do not correspond to annotated exons of protein-coding genes, mRNAs or ESTs[3]. The complexity of nuclear transcriptomes is much higher — fivefold more transcribed base pairs are detected in nuclear RNA than in cytosolic RNA[3], and approximately 80% corresponds to the unannotated portion of the genome[3]. In total, ~15% of all interrogated base pairs can be detected as RNA molecules (either in the cytosol or in the nucleus) in a single human cell line. This is in contrast to a total of 1–2% of base pairs that correspond to the exons of all the annotated protein-coding genes[3]. These data strongly indicate that a significant portion of the human genome can be transcribed. Estimates made by the Encyclopedia Of DNA Elements (ENCODE) consortium[4] — a large multidisciplinary and collaborative effort to characterize the regulatory landscape of ~1% of the human genome — suggest that

Box 1 | **Technologies for mapping RNA expression**

The methods for analysing structure and expression levels of the RNAs described in this Review can be broadly classified into two groups: sequencing-based and hybridization-based approaches.

Sequencing-based approaches. These approaches rely on obtaining direct information about the order of nucleotides in an RNA molecule. They can be further subdivided into methods that involve sequencing of full-length or nearly full-length RNAs, or sequencing of short portions of RNAs, typically derived from the 3′ (SAGE)[113], 5′ (CAGE)[96] or both termini (PET)[97] of the corresponding RNAs. Before sequencing, RNAs are converted into cDNAs that can be further processed to generate truncated cDNAs that contain only short sequences or 'tags' (typically ~14–22 bases) that represent the sequences from either one of the two termini of the original RNA. Generation of tags significantly increases throughput, which in turn significantly increases depth of coverage.

Sequencing-based methods provide the most detailed information about the structure of an RNA molecule, but they have a much lower throughput than hybridization-based methods. Therefore, full-length cDNA sequencing is typically used to catalogue exemplars of different RNA molecules, rather than as a means to comprehensively count the number of molecules in a sample. However, sequencing of short tags of cDNAs, such as CAGE, SAGE and PET tags, has greatly benefited from the increases in throughput and parallelism of sequence-readout methods, and is now used to identify the 5′ and 3′ termini of RNA molecules and estimate their abundance.

Hybridization-based approaches. These methods rely on measuring the magnitude of hybridization of a probe to its target in a complex background, relative to the signal from the background or from control probes. In hybridization-based detection, a probe would detect all molecules that contain regions of complementarity to that probe. If RNA molecules are not separated (using traditional gel-based techniques, that is, northern blots) before hybridization, the net sum of the hybridization signal is the sum of the signal from all the different molecules that can hybridize to a probe.

Compared with the sequencing-based methods, the main advantages of hybridization-based methods are higher throughput and depth of sampling. Throughput is increased by the absence of the requirement for the molecular separation and library construction. The depth of sampling is a key difference between the two types of method — sequencing-based methods provide information about one RNA molecule; although they have benefited from increases in parallelism and throughput, owing to practical limitations, they can provide information about only 10^5–10^7 RNA molecules in each sample, equivalent to the RNA content of 1–30 cells. Hybridization-based methods are intrinsically able to interrogate all RNA molecules in a given sample.

Technology innovations. Both types of method have seen technical innovations to increase parallelism and throughput of the technologies. Sequencing of short pieces of nucleic acids has been most susceptible to the increase in parallelism owing to the advent of such technologies as pyrosequencing, massively parallel signature sequencing (MPSS) and others[114]. The increased parallelisms in hybridization-based technologies largely stems from the advent of high-density microarrays, in which large numbers of probes (as many as 10^5–10^6 probes for different sequences) can be spotted or synthesized on surfaces as small as a square inch[115–117].

Continuing challenges. Most of what is currently known about the sequences of human transcripts is limited to RNA species that are both polyadenylated and long (more than 200 nucleotides). This bias results from technical issues, such as the ease with which polyadenylated RNAs can be purified, and from the conceptual bias that stems from the idea that non-polyadenylated RNAs are unlikely to be functional. Recent studies that use tiling arrays[3] together with earlier studies that used association kinetics[83,84] suggest that the population of non-polyadenylated transcripts is vast, and its complexity exceeds that of the polyadenylated RNAs. Short RNAs represent another heavily underappreciated component of the transcriptome. On the basis of sequencing of libraries of short RNAs[100–108] and tiling array profiling of short RNA populations[9], it seems that the short RNA transcriptome is probably at least as complex, if not more so, than that of the long RNAs. Compartmentalized RNAs are yet another relatively unexplored domain of the transcriptome. Most genome-wide surveys of RNAs have been limited to RNA populations isolated from whole-cell extracts or the cytosolic RNA fraction. Recent results from tiling array profiling of the nuclear transcriptome indicate that this specific subcellular compartment, which accounts for 15–25% of the total cellular RNA, contains an RNA population that is five times more complex than the cytosolic population; most of these RNAs remain to be sequenced[3].

this is in fact the case. Depending on which empirical data sets are included in the estimate, as much as 93% of the genomic sequences in the surveyed ENCODE regions seem capable of being transcribed[5]. This estimate is derived from the union of all intronic and exonic sequences detected by several empirical RNA-mapping technologies in multiple biological samples. A surprisingly large number of unannotated transcripts[6,7] or novel isoforms of protein-coding genes[5] for which primary structures have been elucidated by sequencing do not seem to encode proteins. These transcripts are often referred to as non-coding RNAs (ncRNAs). This is a putative designation, as it is possible that some might in fact encode short proteins or peptides. The term 'transcript of unknown function' (TUF)[3,8] has been proposed by the ENCODE consortium to denote such putative non-coding molecules — thus reserving the label 'ncRNA' for those RNAs for which there is some functional evidence.

Multifunctional usage of the same genomic space is common. Overlapping transcripts can be produced from the same or opposite strands of DNA. The regions of overlap of transcripts from opposite strands can include the exons that are present in mature RNAs, or be mostly confined to the introns. This is exemplified by the phosphatidylserine decarboxylase (*PISD*) gene, which has at least nine overlapping independent transcripts in its genic boundary (FIG. 1). Many detected transcripts contain both exonic and intronic portions of *PISD*, and the 5′ termini of several of these overlapping transcripts are positioned proximal to an empirically determined MYC binding site. Additionally, the same genomic sequences can be shared by both long and short RNAs, suggesting that the function of some of the overlapping coding and non-coding transcripts is to produce short RNAs[9] (see below).

Overlapping transcripts that are made from the same DNA strand can either be functional isoforms (for example, produced by alternative splicing or processing) or lack apparent common functional characteristics (for example, protein-coding potential), despite sharing the same genomic space.

Antisense transcription. Given the extent of transcriptional overlap, it follows that much of it is antisense to protein-coding loci. FIGURE 1 illustrates this for the *PISD* gene, which has nearly as many distinct antisense as sense transcripts. Estimates of the extent of overlapping sense–antisense transcription across the whole genome vary; the highest so far comes from sequence analysis of 158,807 full-length mouse cDNA clones[10]. In this case, the authors reported antisense transcription for 72% of all transcriptional units, with 18,021 (87%) of protein-coding and 13,401 (58.7%) of non-coding transcriptional units having an antisense transcript. Large-scale sequencing of libraries of short sequence tags near the 3′ ends of RNAs (LongSAGE tags) revealed that 9,804 human mRNAs contain an antisense transcript[11]. Analysis of a randomly selected subset of transcripts detected using RACE/tiling arrays found that 61% of all human transcribed regions have

a counterpart on the opposite strand[3]. Furthermore, a recent study showed that sense–antisense pairing is also prevalent in other eukaryotic species, with many sense–antisense pairs being conserved in evolution[12].

Table 1 | **Evidence for widespread transcription**

Method	Organism	Refs
Abundance of polysome-associated polyA RNA	*Homo sapiens*	82
Nucleic acid hybridization reassociation kinetics (Cot curves)	*Strongylocentrotus purpuratus* (sea urchin)	83
	H. sapiens	84
Transcription of satellite DNA on chromosomes in oogenesis	*Triturus cristatus carnifex* (crested newt)	85
Abundance of 5′-capped RNAs versus those with 3′ polyA	*Cricetulus griseus* (hamster)	86
Tiling arrays		
Whole genome	*Escherichia coli*	87
	Arabidopsis thaliana	88
	Drosophila melanogaster	15,89
	H. sapiens	90
	Oryza sativa (rice)	91
	S. purpuratus	92
Whole genome, high resolution	*Saccharomyces cerevisiae*	93
	H. sapiens	9
Ten chromosomes, high resolution	*H. sapiens*	3
Chromosome 21–22	*H. sapiens*	94
Chromosome 22	*H. sapiens*	95
Sequencing		
CAGE tags	*H. sapiens and Mus musculus*	7
CAGE tags to identify promoters	*H. sapiens and M. musculus*	13,96
PET tags	*M. musculus*	97
SAGE tags	*H. sapiens*	98
LongSAGE tags	*H. sapiens*	11,99
Short RNAs	*Caenorhabditis elegans*	100–102
	A. thaliana	103
Testes-specific short RNAs	*H. sapiens, M. musculus and Rattus norvegicus*	104–108
MPSS	*H. sapiens*	109
Human Invitational, 41,118 full-length cDNAs	*H. sapiens*	6
FANTOM2, 60,770 full-length cDNAs	*M. musculus*	110
FANTOM3, 102,281 full-length cDNAs	*M. musculus*	7
Chromatin IP		
ChIP–chip: p53, Sp1, cMyc	*H. sapiens*	23
ChIP–chip: NFκB	*H. sapiens*	24
ChIP–chip: RNA Pol II	*H. sapiens*	111
PET-sequencing-based: p53	*H. sapiens*	112

CAGE, cap analysis of gene expression; ChIP–chip, chromatin immunoprecipitation and hybridization to DNA microarray; FANTOM, Functional Annotation of Mouse; MPSS, massively parallel signature sequencing; PET, paired-end ditag; polyA, polyadenylated; SAGE, serial analysis of gene expression.

Overlapping, same-strand transcription. cDNA sequencing has also revealed significant complexity in transcripts that overlap on the same strand (FIG. 1). Analysis of mouse full-length cDNA clones showed that, on average, 7.6 transcripts can be grouped into a single transcriptional unit on the basis of overlapping genomic location and shared orientation[7]. Furthermore, the majority of transcriptional units (65%) are alternatively spliced[7]. Sequencing of more than 12 million short sequence tags that represented the 5′ termini of RNA polymerase II transcripts (CAGE tags) from multiple human and mouse RNA samples revealed a propensity for transcriptional start sites (TSSs) to mark internal exons and 3′ UTRs of genes, suggesting that there are multiple additional internal initiation sites within the loci of known genes[13].

Long-range interconnected transcription. A further level of genomic architectural complexity is revealed by the recently observed long-range interactions in the genome. Mapping the 5′ ends of 399 genes that lie in the regions chosen by ENCODE has revealed that many use alternative 5′ ends that lie tens and hundreds of kilobases away from the annotated 5′ ends[5,14]. Often, other genes are located between the body of a gene and the distal, previously unannotated, 5′ ends, suggesting that a primary transcript that is initiated at a distal 5′ end must traverse intervening loci on both strands before reaching exons to which it is connected in the mature transcript.

Similar results were observed for *Drosophila melanogaster*, in which distal 5′ ends were found to be on average 20,360 bp from the annotated 5′ ends, and sometimes as far away as 135,000 bp[15]. For comparison, the average length of an annotated intron in a *D. melanogaster* gene is 1,158 bp[15]. The biological importance of these distal regulatory elements is underscored by the observation that many of the distal 5′ ends in *D. melanogaster* transcripts map to the sites of *P*-element insertions that result in well documented fly phenotypes. Because of the genomic distances involved, the effects of many of these *P*-element insertions were previously unconnected with the related distal gene loci, despite the often dramatic phenotypes.

As a consequence of such long-range transcription, multiple exons from previously characterized, separate protein-coding genes can be joined, creating novel spliced transcripts. An example of this is provided by transcripts that originate at a distal 5′ end and join previously unannotated exons with those of caveolin 1 (*CAV1*) and caveolin 2 (*CAV2*) genes (FIG. 2). Recent reports show that such transcriptional fusions of neighbouring genes are surprisingly common in the human genome[16–18]. One of the main implications of these findings is that exons that have been considered to be discrete modules of a specific gene, or at least a genomic locus, must now be considered as more general functional modules that can be joined together in multiple RNA molecules.

Although many of the observed fusion transcripts contain appreciable ORFs, it remains to be determined whether such molecules are translated or whether they function in some as yet unknown fashion. For example, distal 5′ ends and transcriptional fusions might allow

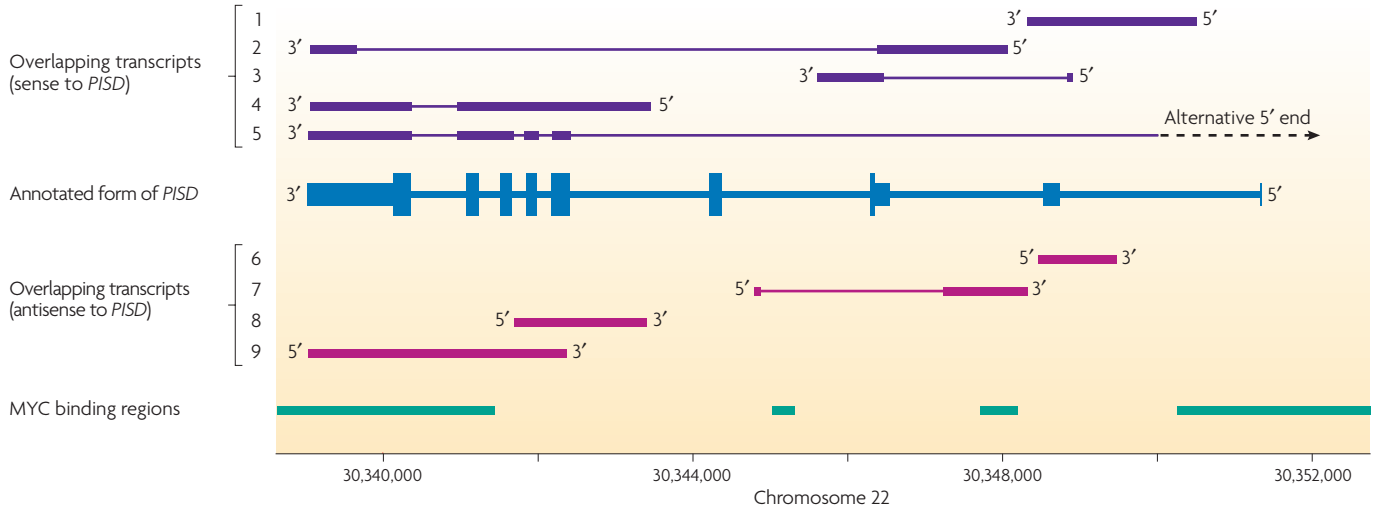

Figure 1 | **Overlapping transcriptional architecture — the *PISD* example.** Five transcripts (in purple) that overlap the RefSeq-annotated form (in blue) of the phosphatidylserine decarboxylase gene (*PISD*) on the same strand, and four transcripts (in pink) that overlap the gene on the opposite strand are shown. The overlapping transcripts were characterized using a combination of RACE and tiling arrays[16]. Binding sites of the transcriptional factor MYC (in green) were determined using a ChIP–chip assay[23]. The coordinates are taken from the hs.NCBIv35 version of the genome.

the combinatorial usage of novel promoters and regulatory DNA regions, or provide intronic RNA, which can be subsequently processed into regulatory RNAs (see discussion below). So, genomic proximity is not always required for exons of different genes to be incorporated into the same RNA molecules.

The mechanisms that generate these recently described fusion transcripts are unclear. Candidate mechanisms include processing of long primary transcripts or splicing between different RNA species (*trans*-splicing). *Trans*-splicing is thought to be less common than transcriptional gene fusion, although naturally occurring examples of *trans*-splicing have been documented for several mammalian genes[19,20]. In some instances, the degree of *trans*-splicing can be high. For example, cloning and sequencing of spliced *MYC* isoforms from a human cell line that overexpresses this gene revealed *trans*-splicing to 33 different genes on 14 different chromosomes[21]. The level of *trans*-splicing can also be high during adenovirus infections, indicating that the virus might overwhelm the endogenous *trans*-splicing regulatory mechanisms[22]. Overall, although the frequency of *trans*-splicing to each RNA molecule is low, it is more readily detectable in highly expressed transcripts. As with gene fusions, it is not clear whether these *trans*-spliced RNA molecules represent functional entities, or whether they are by-products of processes such as transcription of separate genes that lie in close proximity to each other, resulting in occasional *trans*-splicing between their nascent transcripts[19].

Widespread occurrence of promoter regions. A high degree of overlapping transcription implies the existence of regulatory regions other than the canonical locations at the 5′ ends of annotated genes. Unbiased efforts to map transcription factor binding sites found that only 22% of MYC and SP1 binding sites lie proximal to the 5′ ends of well annotated genes, whereas 36% lie within gene boundaries[23]. Another 24% map to intergenic regions. A significant fraction of the sites that were found to lie proximal to the 3′ exons, and that were present in regions that are annotated as internal exons or introns of protein-coding genes, were also seen in the analysis of CAGE tag data[13]. Comparable results were obtained for another transcription factor, NFκB; approximately 28% of NFκB binding sites were found at the 5′ ends of genes and 22% were found more than 50 kb away from known genes[24]. In addition, the same genomic sequences were observed to host overlapping promoters that regulate divergent transcripts in the mammalian genome. Consistent with these recent observations, 10% of all human genes lie 'head-to-head' (that is, they are transcribed in opposite directions, with their promoters being closest to each other), less than 1,000 bp apart, and are regulated by the same genomic sequence[25]. In many such cases, the sequence elements that regulate the divergent genes are shared[25].

The architecture of the eukaryotic transcriptome is clearly much more complex than could have been anticipated in terms of the number of nucleotides that are transcribed and the final arrangements of nucleotides that are present in mature processed RNA molecules. This complexity makes one reconsider the current linear model of genomic organization, and ask what possible advantages such an interleaved genomic organization might offer.

Implications and advantages

Compactly organized and highly interleaved genomes have typically been associated with viruses and microorganisms, for which genome size is limited by the size of the viral particle or cell[26,27]. Such constraints are not known to operate on eukaryotic genomes, which are many orders of magnitude larger. Given the potential problems that are presented by use of the same genomic space for multiple purposes in megabase and gigabase genomes,

SAGE
Serial analysis of gene expression; a technique for mapping the 3′ ends of transcripts.

PET
Paired-end ditag; a method that extracts 36-bp signatures with 18 bp from the 5′ end and another 18 bp from the 3′ end of each cDNA.

Pyrosequencing
A method for DNA sequencing in which the inorganic pyrophosphate (PPi) that is released from a nucleoside triphosphate on DNA chain elongation is detected by a bioluminometric assay.

Massively parallel signature sequencing
A sequencing procedure that allows the reading, in parallel, of short sequence segments of about 17 or 12 nucleotides long, from hundreds of thousands of microbead-attached cDNAs.

LongSAGE
Long serial analysis of of gene expression; a method that allows for the cloning of 20-nucleotide SAGE tags.

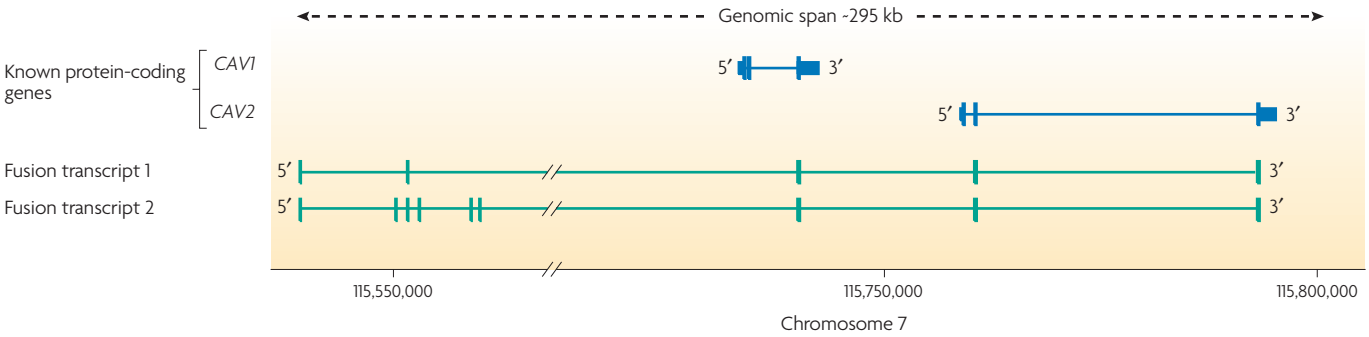

Figure 2 | **Fusion transcripts combining exons of different genes and unannotated regions.** Two different transcripts combine novel 5′ exons with selected exons of caveolin 1 (*CAV1*) and caveolin 2 (*CAV2*). The exons of the two fusion transcripts (GenBank accession numbers EF179101 and EF179102) and *CAV1* and *CAV2* mRNAs are shown as vertical lines; introns are represented as horizontal lines; slanted lines indicate a gap of ~200 kb, used to simplify the depiction of this genomic region. The coordinates are taken from the hs.NCBIv35 version of the genome.

what are the implications and consequences of such a complex genomic organization for a eukaryotic cell?

Increasing protein-coding transcript diversity. One obvious benefit of sharing DNA sequence among different transcripts is the production of diverse protein species from relatively few protein-coding domains (exons). The most prevalent mechanisms to generate mRNA diversity are alternative splicing, alternative initiation of transcription, alternative polyadenylation, gene fusions and *trans*-splicing. The first three processes are common in the higher eukaryotes that have been studied[7,13,28,29]. Analysis of the available genomic annotations, including ESTs, indicates that at least 40–65% of mammalian protein-coding genes could be alternatively-spliced[7,28,29], with ~70% of splicing events occurring in the coding sequence of mRNA[29]. In the mouse, 58% of protein-coding transcriptional units use two or more alternative promoters[13]. Because genome annotation databases are likely to miss information about protein-coding genes that are not highly expressed, these estimates are bound to be conservative. An analysis of 399 human protein-coding loci within ENCODE[4] genomic regions indicated that 90% have either a previously unannotated exon or a new TSS[5,14].

Taken together, these different mechanisms significantly increase the diversity of both transcripts and proteins by using a modular approach to build novel protein and non-protein transcripts from a collection of distal and even non-linear constituents.

Using RNA transcripts as regulatory agents. The potential of RNA as a regulatory molecule — because of its ability to reversibly bind to virtually any other RNA or DNA molecule by nucleotide complementarity — was recognized decades ago[1,30]. Since then, a significant number of RNA-based regulatory systems have been characterized, and many classes of RNA regulators have been discovered in species ranging from viruses to mammals[8,31–36]. RNA has been implicated in the control of RNA stability, gene expression, tissue and cellular development, RNA modification, chromatin organization, alternative splicing, subcellular localization of proteins, heat shock sensing and other processes[8,32–37].

Functional ncRNAs range in size from ~22 bp miRNAs to ~18 kb *XIST* (X-inactive-specific transcript) and ~108 kb *AIR* (antisense *IGF2R* RNA) ncRNAs. Such tremendous size variation, coupled with the growing realization that long ncRNAs might exceed protein-coding mRNAs in number if not functional diversity, highlight the underappreciated importance of ncRNAs in the cell (reviewed in REFS 8,38). Below we discuss the possible advantages of using RNA-based regulatory systems, of the sort envisioned by Mattick[38–41], in the context of interleaved genomic organization.

A key element in using RNA as a regulatory agent is its full or partial sequence complementarity to the target[32,35,42,43]. One way to facilitate this is to transcribe multiple overlapping independent RNAs that contain the same genomic regions but in the context of different transcripts. As such, the intended interactions might be more likely to occur because they involve *cis* (caused by production of RNAs from the opposite strands of the same DNA sequence) rather than *trans* (caused by interactions between RNAs that are produced from different regions of genome) events. The obvious advantages of *cis*-based RNA signalling is the localization of interacting RNA components, which could also simplify the task of localizing various participating protein components to the site of the RNA–RNA or RNA–DNA interactions (for example, mediation of target RNA transcription and subsequent stability or regulation of its interaction with ribosomes)[33–35]. The utilization of RNAs or portions of RNAs as *trans*-targets is not precluded by such a strategy, and such interactions would be expected to evolve to increase the overall efficiency of such an approach (see FIG. 3 for a hypothetical model).

So far, the known effectors of *trans*-RNA-based signalling have been mostly limited to RNAs that are <200 nucleotides long, and that can be referred to as short RNAs. Two prominent classes of regulatory short RNAs, microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs), are produced from longer precursor RNA molecules[32,42,44,45]. A sizeable fraction of known miRNAs and almost all snoRNAs are found within annotated genes or non-coding transcriptional units[45–47]. In many cases, these

RACE/tiling arrays
An unbiased, high-throughput method to identify the extents of DNA products from rapid amplification of cDNA ends (RACE) reactions by hybridizing them to tiling arrays.

CAGE
Cap analysis of gene expression; a technique for mapping the 5′ ends of transcripts.

P element
A member of a family of transposable elements that are widely used as the basis of tools for mutating and manipulating the genome of *Drosophila melanogaster*.

ChIP–chip
A method that combines chromatin immunoprecipitation with microarray technology to identify *in vivo* sites of protein–DNA interactions.

MicroRNA
A form of ssRNA, typically 20–25 nucleotides long that is thought to regulate the expression of other genes, either through inhibiting protein translation or degrading a target mRNA transcript through a process that is similar to RNAi.

snoRNA
A type of small RNA, the functions of which include RNA cleavage and specification of sites of ribose methylation and pseudouridylation.

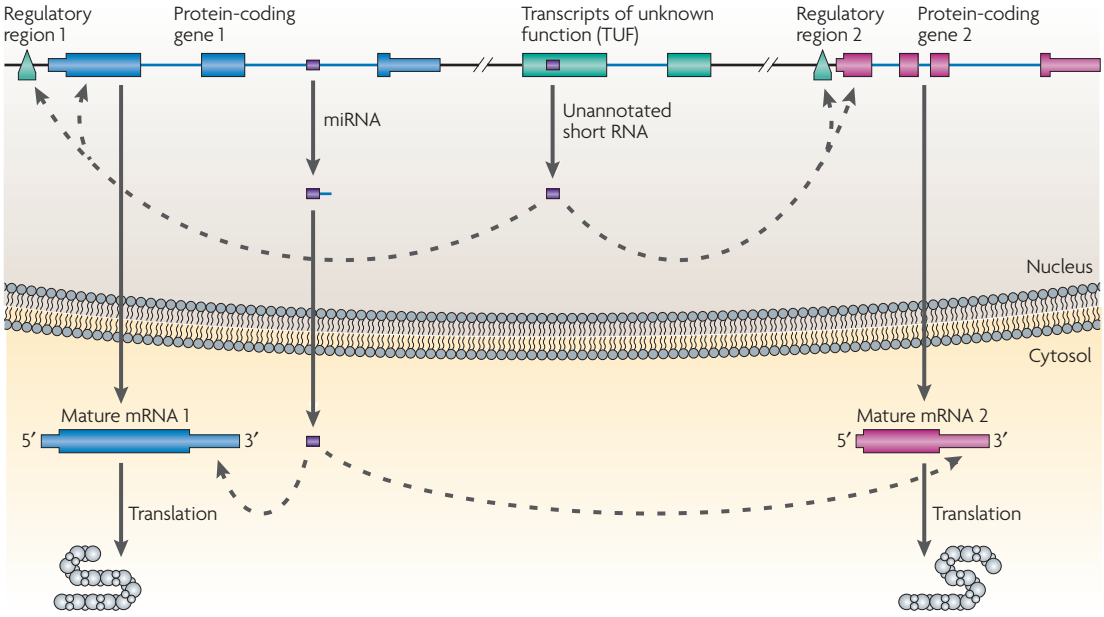

Figure 3 | **RNA-based signalling pathways.** A microRNA (miRNA) is encoded by DNA sequences that lie within an intron (blue) of protein-coding gene 1. Expression of the miRNA is coupled to expression of the 'host' gene. The miRNA is processed from a long precursor RNA, which contains the intronic sequences, in the nucleus and then fully matures in the cytosol. The mature miRNA is believed to influence expression and stability of other mRNAs (potentially including that of its own host gene) in *trans* (indicated by broken arrows in the cytosol), through partial nucleotide complementarity with the 3′ UTR sequences of target mRNAs. A similar strategy is outlined for a hypothetical unannotated short RNA (miRNA or other short RNA) that is encoded by a transcript of unknown function. A large amount of previously unannotated stable short RNAs has been discovered in mammals[9,103–107], *Caenorhabditis elegans*[100,101] and *Arabidopsis thaliana*[103]. By analogy with known short RNAs, other short RNAs probably act by regulating gene expression in *cis* or *trans* modes (*trans* modes are indicated by broken arrows in the nucleus). Novel short RNAs could target regulatory regions or parts of mRNAs. Other short RNAs could potentially be found within the annotated boundaries of known genes, possibly encoded by overlapping transcripts that are regulated by their own promoters (not shown).

short RNAs are produced from the intronic sequences of these genes or from separate overlapping transcripts that have been cleaved by ribonucleases and associated factors[42,44,45]. Co-expression analysis of miRNAs and their host genes in humans[48] strongly suggested that the first mechanism is likely to be responsible for production of human regulatory short RNAs. However, a comprehensive analysis of the expression of 127 ncRNAs, including snoRNAs and small nuclear RNAs (snRNAs), in *Caenorhabditis elegans* showed that the intronic ncRNAs might form two classes: those that are co-regulated with the expression of the genes in which they lie and those that are apparently regulated by their own promoters[49].

Once it has been processed from a longer transcript, a mature short RNA can act in *cis* or *trans* to regulate other RNA molecules. Two hypothetical scenarios for this mode of regulation, one involving a known miRNA pathway and one including a hypothetical as yet unannotated short RNA, are shown in FIG. 3. The potential for global genome regulation by this RNA-mediated signal is significant, as regulatory RNA–RNA interactions could be effected by short stretches of full or partial complementarity. Each miRNA has been estimated to potentially regulate 200 targets on average, and thousands of human genes could be regulated by RNA–RNA duplexes as short as 7 bp[50–52]. It has also been estimated that as many as 20–30% of human genes are regulated by miRNAs[51,53,54] and an equally large number of genes are under selective pressure to avoid complementarity with miRNAs[55]. The ability to couple the expression of such short RNA regulators to the expression of protein-coding genes by processing them from spliced introns has obvious evolutionary advantages if the RNA regulators and their host genes are involved in the same pathways in a cell, like the snoRNAs and their host genes[39,40,45], or if short RNA transcription *per se* is used as a means of signalling the expression status of the host genes[39,40]. Additional plasticity in the timing of expression could be achieved by expressing the internal short RNA regulators from independently regulated internal promoters. These aspects make both strategies attractive, even in the context of the same locus, so an overlapping transcriptional architecture would be advantageous.

Although several families of short RNA regulators have been identified, an ever-growing number of individual long ncRNAs have also been found to function in various key cellular processes. One of the most rapidly evolving human genes identified to date, highly accelerated

snRNA
A small RNA molecule that functions in the nucleus by guiding the assembly of macromolecular complexes on the target RNA to allow site-specific modifications or processing reactions to occur.

region 1A (*HAR1A*), is transcribed into an ncRNA that is expressed during brain cortical development[56]. Long putative ncRNAs have been found as markers for hepatocellular carcinomas[57] and cell fate during mammary gland development[58]. Long ncRNAs have also been directly linked to regulation of transcription factors such as calcium-sensing nuclear factor of activated T-cells (NFAT)[59] and the fly homeotic gene *Ultrabithorax* (*Ubx*) (through an interaction with the histone methyltransferase ASH1)[60].

Using transcription as a regulatory process. Overlapping transcription could also indicate a different type of regulation that does not rely on RNA-mediated signalling, but rather on the act of transcription *per se*. It is has been widely proposed that passage of the RNA pol II complex through a region of DNA might remove the histones, thereby resetting its chromatin structure. Consistent with this model, long-range transcription has been observed at several genomic loci. Many discrete DNA elements controlling the expression of genes that are separated from them by tens or hundreds of kilobases have been identified in eukaryotic genomes[61–64]. Such long-range elements can enhance or suppress expression of one or more genes. Elements that increase gene expression can be separated into enhancers and locus control regions (LCRs), whereas elements that suppress gene expression can be classed as silencers or elements that control the imprinted status of a locus (imprint control centres). Transcription has a role in the functional activities of each of these long-range elements.

Long-range DNA elements have been found in association with ncRNAs[60,65–67]. Some of these ncRNAs exceed 100 kb in length and span multiple genes[68]. Effects of non-coding transcripts on the expression of overlapping and neighbouring genes were directly investigated in four human and mouse loci: β-globin, *CD79b-GH*, the *KCNQ1* cluster and the *IGF2R* cluster[69–72]. When DNA elements that cause transcriptional termination were inserted near the promoter elements of these ncRNAs, long-range transcription was abrogated leading to either loss of imprinting (*KCNQ1* and *IGF2R* clusters) or decreased gene expression for loci that were positively regulated by a distal LCR (*GH*) or enhancer element (β-globin).

These studies show that ncRNA transcription is a major regulatory mechanism of long-range control elements. However, they do not distinguish between the importance of transcription *per se* and the production of functional RNA products. The model of transcription as a linear, 'snow-plow' regulator is challenged by the observation that, in the case of the two imprinted clusters *KCNQ1* and *IGF2R*, genes that did not directly overlap the ncRNAs were equally as affected as those that did. Therefore, it has been suggested that the ncRNA transcription could alter the state of other long-range DNA elements, which could subsequently silence the genes in the imprinted clusters that do not directly overlap with the long ncRNAs[73].

Co-regulation of different genes by the same DNA elements might be yet another example of the advantages of the interleaved genomic organization. Analysis of human head-to-head genes revealed that expression levels were correlated across many published microarray datasets, with a high degree of statistical significance[25]. Interestingly, functional analysis of ten randomly selected bidirectional promoters showed that most contain shared elements that regulate expression from both strands, rather than housing two non-overlapping promoters that function in opposite directions[25]. Additionally, one promoter sequence can regulate more than one downstream gene, as suggested by the prevalence of transcriptional gene fusions in the human genome[16–18].

Evolutionary implications

Implications at the macro-scale: global genomic organization. If the interleaved genomic organization is evolutionarily advantageous, one would expect to see conservation of genomic architecture and transcript organization (for example, that the same two transcripts would overlap or that internal promoters would overlap in syntenic exons of orthologous genes in more than one species).

The data to support this come from analysis of the conservation of patterns of overlapping genes and transcripts in prokaryotes and eukaryotes. Perhaps the most unexpected observation comes from the microbial genomes, in which overlapping transcription is relatively common[26,27]. In these cases, the frequency of overlapping transcripts does not seem to correlate with genome compactness as measured by the distance between the genes or by the proportion of coding sequence in a genome[74]. Interestingly, microbial genes that overlap tend to have more orthologues in other species and therefore seem to be more evolutionarily conserved[74], yet, surprisingly, 70% overlap by less than 15 bp[74].

Although the sense–antisense overlap arrangement seems to be less common than the sense–sense overlap in microbial genomes, the first tends to be longer, suggesting a different type of a relationship than in the case of genes that overlap on the same strand, perhaps arising from sense–antisense RNA complementarity[74]. Together, these observations indicate that the overlapping arrangement of genes could have potential regulatory roles beyond merely helping to reduce genome size[74].

The patterns of overlapping genes seem to be conserved in mammals. Analysis of ten sequenced animal genomes showed that, of the 3,915 human sense–antisense pairs, at least 313 are preserved in the mouse[12]; several human sense–antisense pairs are also conserved in evolutionarily distant organisms such as frogs[12].

The reasons for conserving such sense–antisense pairing could involve gene regulation. Indeed, expression analysis in 16 human tissues showed that genes in such pairs tend to either be co-expressed or have an inverse correlation in levels of expression[75]. The same study also showed that sense–antisense pairs that fit these expression patterns tend to be conserved between humans and mice[75]. Interestingly, a study of a pair of yeast genes, *GAL10* and *GAL7*, revealed a drastic decrease in gene expression caused by converging transcription complexes *in vivo*[76]. However, an inverse correlation of

Locus control region
A *cis*-acting sequence that organizes a gene cluster into an active chromatin block and enhances transcription.

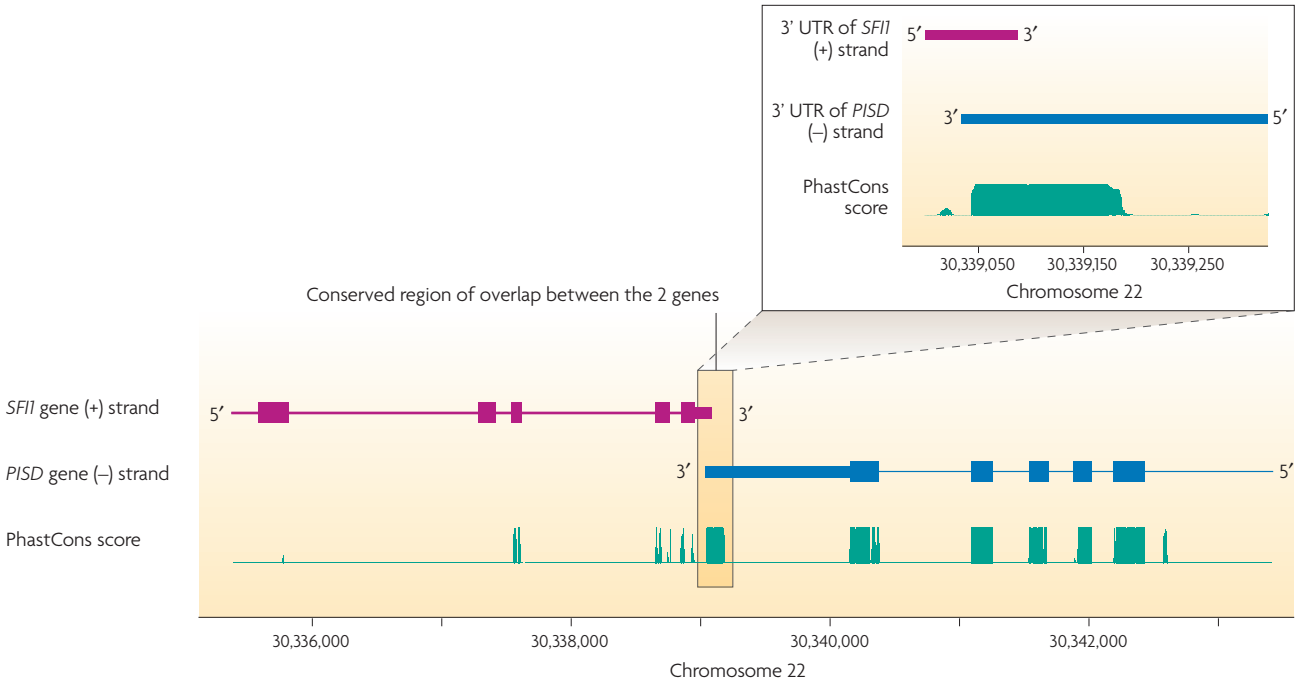

Figure 4 | **Conservation of an overlapping region of two genes on opposite strands.** Conservation of the overlapping region of spindle assembly associated *sfi1* homologue (*SFI1*) and phosphatidylserine decarboxylase (*PISD*) is as high as that of the protein-coding exons (represented by wider boxes). The untranslated portions of each mRNA are presented as narrower boxes. The conservation score of each nucleotide (PhastCons score) is represented by Vertebrate Conservation scores on the scale of 0.2–1.0. The annotations and conservation scores were loaded from the UCSC Genome Browser (University of California at Santa Cruz) [118,119]. The coordinates are taken from the hs.NCBIv35 version of the genome.

sense–antisense expression levels is not universal, as several other reports show positive correlation between levels of sense and antisense transcripts[10,23,77,78]. Thus, use of the same region of DNA as a template for overlapping transcription does not seem to predetermine an expression relationship.

Implications at the micro-scale: nucleotide level. Although interleaved transcriptional architecture occurs at some loci and is conserved between distant species, it represents an added constraint on the ability of a genome to evolve in the region of overlap. It is generally assumed that the rate at which a base changes or a region of DNA accepts insertions or deletions is inversely related to the importance of its biological function. The degree of base-pair conservation is expected to be proportional to the number of functional elements that use it. An example of this is provided by higher conservation scores in the overlapping region of the 3′ UTRs of the human genes spindle assembly associated *sfi1* homologue (*SFI1*) and *PISD*, which are located on the opposite strands of the genome (FIG. 4).

Sequence conservation is higher at and around the overlapping region compared with the intronic and UTR sequences that do not overlap, and is as conserved as the protein-coding portions of the mRNAs (FIG. 4). Furthermore, a whole-genome analysis of the sites of short- and long-RNA production in human tissue culture cells indicates that nucleotides that are shared by long and short RNAs tend to be more conserved than those that are shared by only long RNAs[9]. Influencing this is the overall fitness of an organism, which might depend on how essential the function of each overlapping element is. If a region or a base pair has more than one function, all of these might be affected by a single nucleotide change.

Perhaps the most apparent example of this effect is the case of overlapping ORFs, which are common in bacterial and viral genomes. In such cases, a base pair change affects both ORFs, which explains why the rates of change in the region of overlap tend to be reduced[26,27,79]. However, the degree to which a nucleotide change affects any given function might be different. For example, a nucleotide substitution could be synonymous in one ORF and non-synonymous in an overlapping ORF[26]. In cases in which different types of functional elements overlap (for example, an exon and a promoter), a nucleotide substitution could bring about an amino-acid change if it occurs in the first two positions of a codon, and change promoter strength if it occurs in a crucial promoter element.

Implications for the interpretation of nucleotide changes. Mutations at non-annotated genomic sites, such as intronic regions that are distal from splice sites, can affect fitness if they occur at internal promoter regions, in an exon of an overlapping transcript, or in a short RNA. Thus, a phenotype that is associated with a DNA

sequence change could be a sum of the phenotypes caused by the change in all elements that share this sequence. Furthermore, if the overlapping transcripts have non-redundant functions, the phenotypic effect of a sequence change in a shared exon could manifest itself more severely than a change in an exon that is unique to only one such isoform. The magnitude of the phenotypic effect of such a mutation might not only be proportional to a direct biochemical effect of that change in any one element, but also to the combined effects of disruption of numerous overlapping transcripts. Moreover, a mutation that does not seem to affect the function of one element (that is, synonymous substitutions in protein-coding regions) might affect other elements that share that sequence.

A mutation can affect a gene that has been annotated as distal or that is separated from the site of mutation by intervening genes if the mutation occurs at an alternative 5′ end of that gene. For example, a significant fraction of unannotated transcripts in the *D. melanogaster* transcriptome might be identified as unannotated extensions of known protein-coding genes[15]. A number of these distal 5′ ends map to the sites of lethal mutations caused by *P*-element insertions. Genetic complementation crosses between the fly lines that contained these distal 5′ *P*-element insertions and those that carried transposon insertions in annotated exons of the corresponding genes confirmed the functional relationship between such distal 5′ ends and the genes to which they were connected[15]. The discovery of many unannotated and distal TSSs suggests that a similar mechanism might also occur in human disease, raising the possibility that many more SNPs could be associated with human disorders. Overall, it is not uncommon for sequence polymorphisms that lie in 'non-coding' regions to be associated with certain disease conditions. Interpretation of such SNPs is complicated by the fact that our knowledge of the genome and its organization is far from complete.

The complex phenotypic implications of the interleaved genomic architecture are highlighted by the example of a SNP that is associated with a predisposition to an autoimmune thyroid disease; the SNP lies in an intron of *ZFAT*, which encodes a zinc-finger protein[80]. The SNP coincides with the 3′ UTR of a truncated form of *ZFAT* (*TR-ZFAT*) and the promoter of an overlapping transcript, (*SAS-ZFAT*), which lies in an antisense orientation to *TR-ZFAT*. The RNA levels of *TR-ZFAT* are positively correlated with the SNP, whereas the levels of *SAS-ZFAT* are negatively correlated. However, the SNP does not directly affect the stability of *TR-ZFAT* mRNA; instead, the levels of *TR-ZFAT* RNA seem to be downregulated by the *SAS-ZFAT* antisense transcript[80]. Therefore, the SNP probably influences expression of an isoform of the gene that carries the intron in which it resides, by influencing expression of an antisense transcript[80].

Evolutionary conservation scores
A quantitative measure of evolutionary relationships derived from comparative analysis of genomic DNA sequences from multiple species. Phastcons are one type of evolutionary conservation score.

Conclusions

The continuing pace of the discovery of novel transcripts and regulatory regions strongly suggests that regions of the genome that are considered 'non-coding' or 'non-functional' might be alternatively regarded as 'currently unannotated'. However, the functional annotation of a genome will not be limited to charting the function of the unannotated genomic space; it will also provide new functions for those regions that are already annotated. The final organization of transcribed nucleotides into mature RNAs represents another level of complexity that has been significantly underappreciated until now. The purposes of interleaved transcription could include combinatorial usage of DNA regulatory regions, increasing protein diversity, clearing DNA of existing chromatin marks, facilitating regulation of gene expression through RNA–RNA or RNA–DNA interactions, and creating long RNAs that either function themselves as long ncRNAs or serve as the progenitors of short RNAs.

An interleaved genomic organization poses important mechanistic challenges for the cell. One involves the steric issues that stem from using the same DNA molecules for multiple functions. The overlap of functionally important sequence motifs must be resolved in time and space for this organization to work properly. Another challenge is the need to compartmentalize RNA or mask RNAs that could potentially form long double-stranded regions, to prevent RNA–RNA interactions that could prompt apoptosis. Despite these challenges, the existence of this interleaved genomic organization seems to have clear evolutionary advantages. Perhaps the clearest indication of this comes from microbial genomes, in which overlapping gene organization is fairly common but does not correlate with genomic compactness. The functional annotations of the human and other eukaryotic genomes are far from complete; however, initial observations suggest that an interleaved organization is recapitulated in these more complex systems.

A concerted effort to create a catalogue of a diverse variety of functional elements has been initiated by the National Human Genome Research Institute (NHGRI), USA, under the auspices of the ENCODE project[4]. This catalogue includes maps of sites of transcription, TSSs, DNase hypersensitive genomic regions, promoters, transcription regulatory elements, origins of replication and other elements, derived from multiple human samples that were selected from diverse developmental origins or from a time course of differentiation. The availability of various maps of different functional elements, combined with evolutionary conservation scores for individual base pairs[81] (made possible by the availability of several genome sequences), should provide fascinating insights into the interrelationship between the rate of evolutionary change and interleaved genomic organization.

Similar projects to catalogue functional regions in model organisms such as *D.melanogaster* and *C. elegans* have been proposed, to be followed by efforts to decipher the biological roles and importance of the newly identified functional elements. Therefore, the scientific community is now faced with a unique opportunity to consider and organize multidisciplinary programmes to achieve these aims.

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
A seminal work on the regulation of gene expression; the first to suggest that RNA could have a role.
2. Johnson, J. M., Edwards, S., Shoemaker, D. & Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21**, 93–102 (2005).
3. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
4. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306**, 636–640 (2004).
5. ENCODE Project Consortium. The ENCODE pilot project: identification and analysis of functional elements in 1% of the human genome. *Nature* (in the press).
6. Imanishi, T. *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2**, e162 (2004).
7. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
This reference provides an unparalleled insight into the complexity of the mouse transcriptome on the basis of sequencing of full-length cDNAs and cDNA tags.
8. Willingham, A. T. & Gingeras, T. R. TUF love for 'junk' DNA. *Cell* **125**, 1215–1220 (2006).
9. Kapranov, P. *et al.* Genome-wide RNA maps reveal interlaced transcript architecture, new classes of RNAs and possible function for pervasive transcription. *Science* (in the press).
10. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
11. Ge, X., Wu, Q., Jung, Y. C., Chen, J. & Wang, S. M. A large quantity of novel human antisense transcripts detected by LongSAGE. *Bioinformatics* **22**, 2475–2479 (2006).
12. Zhang, Y., Liu, X. S., Liu, Q. R. & Wei, L. Genome-wide *in silico* identification and analysis of *cis* natural antisense transcripts (*cis*-NATs) in ten species. *Nucleic Acids Res.* **34**, 3465–3475 (2006).
13. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38**, 626–635 (2006).
14. Denoeud, F. *et al.* Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* (in the press).
15. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nature Genet.* **38**, 1151–1158 (2006).
16. Kapranov, P. *et al.* Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15**, 987–997 (2005).
17. Parra, G. *et al.* Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.* **16**, 37–44 (2006).
18. Akiva, P. *et al.* Transcription-mediated gene fusion in the human genome. *Genome Res.* **16**, 30–36 (2006).
References 14–18 were the first studies to detail developmental and tissue- or cell-type-specific regulatory regions that are distal from the genes they regulate, often utilizing promoters and exons from upstream genes to form chimeric versions of well annotated protein-coding transcripts.
19. Horiuchi, T. & Aigaki, T. Alternative *trans*-splicing: a novel mode of pre-mRNA processing. *Biol. Cell* **98**, 135–140 (2006).
20. Finta, C., Warner, S. C. & Zaphiropoulos, P. G. Intergenic mRNAs. Minor gene products or tools of diversity? *Histol. Histopathol.* **17**, 677–682 (2002).
21. Chen, C. *et al.* High frequency *trans*-splicing in a cell line producing spliced and polyadenylated RNA polymerase I transcripts from an rDNA–myc chimeric gene. *Nucleic Acids Res.* **33**, 2332–2342 (2005).
22. Kikumori, T., Cote, G. J. & Gagel, R. F. Naturally occurring heterologous *trans*-splicing of adenovirus RNA with host cellular transcripts during infection. *FEBS Lett.* **522**, 41–46 (2002).
23. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
24. Martone, R. *et al.* Distribution of NF-κB-binding sites across human chromosome 22. *Proc. Natl Acad. Sci. USA* **100**, 12247–12252 (2003).
References 23 and 24 represent the first reports of the unbiased profiling of transcription factor binding sites and provide the first comprehensive evidence for the utilization of promoters in non-canonical genomic locations.
25. Trinklein, N. D. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res.* **14**, 62–66 (2004).
26. Krakauer, D. C. Stability and evolution of overlapping genes. *Evolution* **54**, 731–739 (2000).
27. Shcherbakov, D. V. & Garber, M. B. Overlapping genes in bacterial and bacteriophage genomes. *Mol. Biol. (Mosk)* **34**, 572–583 (2000).
28. Sharov, A. A., Dudekula, D. B. & Ko, M. S. Genome-wide assembly and analysis of alternative transcripts in mouse. *Genome Res.* **15**, 748–754 (2005).
29. Zavolan, M. *et al.* Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* **13**, 1290–1300 (2003).
30. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
31. Gupta, A., Gartner, J. J., Sethupathy, P., Hatzigeorgiou, A. G. & Fraser, N. W. Anti-apoptotic function of a microRNA encoded by the HSV-1 latency-associated transcript. *Nature* **442**, 82–85 (2006).
32. Zamore, P. D. & Haley, B. Ribo-gnome: the big world of small RNAs. *Science* **309**, 1519–1524 (2005).
33. Mattick, J. S. & Makunin, I. V. Small regulatory RNAs in mammals. *Hum. Mol. Genet.* **14**, R121–R132 (2005).
34. Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Hum. Mol. Genet.* **15**, R17–R29 (2006).
35. Storz, G., Altuvia, S. & Wassarman, K. M. An abundance of RNA regulators. *Annu. Rev. Biochem.* **74**, 199–217 (2005).
36. Goodrich, J. A. & Kugel, J. F. Non-coding-RNA regulators of RNA polymerase II transcription. *Nature Rev. Mol. Cell Biol.* **7**, 612–616 (2006).
37. Prasanth, K. V. & Spector, D. L. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.* **21**, 11–42 (2007).
A comprehensive review of ncRNAs.
38. Mattick, J. S. RNA regulation: a new genetics? *Nature Rev. Genet.* **5**, 316–323 (2004).
39. Mattick, J. S. Introns: evolution and function. *Curr. Opin. Genet. Dev.* **4**, 823–831 (1994).
40. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* **2**, 986–991 (2001).
41. Mattick, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* **25**, 930–939 (2003).
References 38–41 review the concept of RNA as a carrier of information in the cell.
42. Kim, V. N. & Nam, J. W. Genomics of microRNA. *Trends Genet.* **22**, 165–173 (2006).
43. Kiss, T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109**, 145–148 (2002).
44. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
45. Filipowicz, W. & Pogacic, V. Biogenesis of small nucleolar ribonucleoproteins. *Curr. Opin. Cell Biol.* **14**, 319–327 (2002).
46. Huang, Z. P. *et al.* Genome-wide analyses of two families of snoRNA genes from *Drosophila melanogaster*, demonstrating the extensive utilization of introns for coding of snoRNAs. *RNA* **11**, 1303–1316 (2005).
47. Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. & Bradley, A. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14**, 1902–1910 (2004).
48. Baskerville, S. & Bartel, D. P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**, 241–247 (2005).
49. He, H. *et al.* Profiling *Caenorhabditis elegans* non-coding RNA expression with a combined microarray. *Nucleic Acids Res.* **34**, 2976–2983 (2006).
50. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nature Genet.* **37**, 495–500 (2005).
51. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
52. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).
53. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
54. Lim, L. P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
55. Farh, K. K. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**, 1817–1821 (2005).
56. Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167–172 (2006).
57. Lin, R., Maeda, S., Liu, C., Karin, M. & Edgington, T. S. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* **26**, 851–858 (2006).
58. Ginger, M. R. *et al.* A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc. Natl Acad. Sci. USA* **103**, 5781–5786 (2006).
59. Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570–1573 (2005).
An example of the use of high-throughput technologies to elucidate the function of human ncRNAs.
60. Sanchez-Elsner, T., Gou, D., Kremmer, E. & Sauer, F. Noncoding RNAs of trithorax response elements recruit *Drosophila* ASH1 to Ultrabithorax. *Science* **311**, 1118–1123 (2006).
61. Dean, A. On a chromosome far, far away: LCRs and gene expression. *Trends Genet.* **22**, 38–45 (2006).
62. Li, Q., Peterson, K. R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077–3086 (2002).
63. Lewis, A. & Reik, W. How imprinting centres work. *Cytogenet. Genome Res.* **113**, 81–89 (2006).
64. Zuniga, A. Globalisation reaches gene regulation: the case for vertebrate limb development. *Curr. Opin. Genet. Dev.* **15**, 403–409 (2005).
65. Ling, J., Baibakov, B., Pi, W., Emerson, B. M. & Tuan, D. The HS2 enhancer of the β-globin locus control region initiates synthesis of non-coding, polyadenylated RNAs independent of a *cis*-linked globin promoter. *J. Mol. Biol.* **350**, 883–896 (2005).
66. Masternak, K., Peyraud, N., Krawczyk, M., Barras, E. & Reith, W. Chromatin remodeling and extragenic transcription at the MHC class II locus control region. *Nature Immunol.* **4**, 132–137 (2003).
67. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human β-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).
68. O'Neill, M. J. The influence of non-coding RNAs on allele-specific gene expression in mammals. *Hum. Mol. Genet.* **14**, R113–R120 (2005).
69. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813 (2002).
70. Mancini-Dinardo, D., Steele, S. J., Levorse, J. M., Ingram, R. S. & Tilghman, S. M. Elongation of the *Kcnq1ot1* transcript is required for genomic imprinting of neighboring genes. *Genes Dev.* **20**, 1268–1282 (2006).
71. Ho, Y., Elefant, F., Liebhaber, S. A. & Cooke, N. E. Locus control region transcription plays an active role in long-range gene activation. *Mol. Cell* **23**, 365–675 (2006).
72. Ling, J. *et al.* HS2 enhancer function is blocked by a transcriptional terminator inserted between the enhancer and the promoter. *J. Biol. Chem.* **279**, 51704–51713 (2004).
References 69–72 show that long-range transcription is required for gene activation and silencing.
73. Pauler, F. M. & Barlow, D. P. Imprinting mechanisms — it only takes two. *Genes Dev.* **20**, 1203–1206 (2006).
74. Johnson, Z. I. & Chisholm, S. W. Properties of overlapping genes are conserved across microbial genomes. *Genome Res.* **14**, 2268–2272 (2004).
75. Chen, J., Sun, M., Hurst, L. D., Carmichael, G. G. & Rowley, J. D. Genome-wide analysis of coordinate expression and evolution of human *cis*-encoded sense–antisense transcripts. *Trends Genet.* **21**, 326–329 (2005).
76. Prescott, E. M. & Proudfoot, N. J. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl Acad. Sci. USA* **99**, 8796–8801 (2002).

77. Jen, C. H., Michalopoulos, I., Westhead, D. R. & Meyer, P. Natural antisense transcripts with coding capacity in *Arabidopsis* may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol.* **6**, R51 (2005).
78. Moorwood, K. *et al.* Antisense *WT1* transcription parallels sense mRNA and protein expression in fetal kidney and can elevate protein levels in vitro. *J. Pathol.* **185**, 352–359 (1998).
79. Miyata, T. & Yasunaga, T. Evolution of overlapping genes. *Nature* **272**, 532–535 (1978).
80. Shirasawa, S. *et al.* SNPs in the promoter of a B cell-specific antisense transcript, SAS-ZFAT, determine susceptibility to autoimmune thyroid disease. *Hum. Mol. Genet.* **13**, 2221–2231 (2004).
An example of an intronic SNP that causes predisposition to a disease by influencing the levels of an antisense transcript.
81. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
82. Milcarek, C., Price, R. & Penman, S. The metabolism of a poly(A) minus mRNA fraction in HeLa cells. *Cell* **3**, 1–10 (1974).
83. Hough, B. R., Smith, M. J., Britten, R. J. & Davidson, E. H. Sequence complexity of heterogeneous nuclear RNA in sea urchin embryos. *Cell* **5**, 291–299 (1975).
84. Holland, C. A., Mayrand, S. & Pederson, T. Sequence complexity of nuclear and messenger RNA in HeLa cells. *J. Mol. Biol.* **138**, 755–778 (1980).
85. Varley, J. M., Macgregor, H. C. & Erba, H. P. Satellite DNA is transcribed on lampbrush chromosomes. *Nature* **283**, 686–688 (1980).
86. Saldit-Georgieff, M., Harpold, M. M., Wilson, M. C. & Darnell, J. E. Jr. Large heterogeneous nuclear ribonucleic acid has three times as many 5′ caps as polyadenylic acid segments, and most caps do not enter polyribosomes. *Mol. Cell. Biol.* **1**, 179–187 (1981).
References 82–86 provide the first indications that a large fraction of the eukaryotic genome is transcribed, and that non-polyadenylated RNA is prevalent.
87. Selinger, D. W. *et al.* RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nature Biotechnol.* **18**, 1262–1268 (2000).
88. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).
89. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**, 655–660 (2004).
90. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
91. Li, L. *et al.* Genome-wide transcription analyses in rice using tiling microarrays. *Nature Genet.* **38**, 124–129 (2006).
92. Samanta, M. P. *et al.* The transcriptome of the sea urchin embryo. *Science* **314**, 960–962 (2006).
93. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103**, 5320–5325 (2006).
94. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
The first unbiased high-resolution microarray-based study of the genomics era, showing that the transcriptional complexity of human cytosolic polyadenylated RNA is up to an order of magnitude more complex that can be explained by exons of known genes.
95. Rinn, J. L. *et al.* The transcriptional activity of human Chromosome 22. *Genes Dev.* **17**, 529–540 (2003).
96. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
97. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods* **2**, 105–111 (2005).
98. Chen, J. *et al.* Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl Acad. Sci. USA* **99**, 12257–12262 (2002).
99. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20**, 508–512 (2002).
100. Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13**, 807–818 (2003).
101. Deng, W. *et al.* Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res.* **16**, 20–29 (2006).
102. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
103. Lu, C. *et al.* Elucidation of the small RNA component of the transcriptome. *Science* **309**, 1567–1569 (2005).
104. Aravin, A. *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203–207 (2006).
105. Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199–202 (2006).
106. Grivna, S. T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.* **20**, 1709–1714 (2006).
107. Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313**, 363–367 (2006).
108. Watanabe, T. *et al.* Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev.* **20**, 1732–1743 (2006).
109. Jongeneel, C. V. *et al.* An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res.* **15**, 1007–1014 (2005).
110. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
111. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
112. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).
113. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
114. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 1767–1776 (2005).
115. Elvidge, G. Microarray expression technology: from start to finish. *Pharmacogenomics* **7**, 123–134 (2006).
116. Kapranov, P., Sementchenko, V. I. & Gingeras, T. R. Beyond expression profiling: next generation uses of high density oligonucleotide arrays. *Brief. Funct. Genomic. Proteomic.* **2**, 47–56 (2003).
117. Mockler, T. C. *et al.* Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**, 1–15 (2005).
118. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
119. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

Acknowledgements

We apologize to the authors whose primary work has not been cited due to the space constraints. Some of the work described in this Review has been funded in part with Federal Funds from the US National Cancer Institute and from the US National Human Genome Research Institute, and by Affymetrix. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Service, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

Competing interests statement

The authors declare competing financial interests: see web version for details.

DATABASES

The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
CAV1 | *CAV2* | *GAL7* | *GAL10* | *HAR1A* | *MYC* | *PISD* | *SFI1* | *UBX* | *XIST* | *ZFAT*
UniProtKB: http://ca.expasy.org/sprot
MYC | NFAT | NFκB | SP1

FURTHER INFORMATION

Affymetrix Human Transcriptome web site: http://transcriptome.affymetrix.com
ENCODE: http://www.genome.ucsc.edu/ENCODE
RNAdb database of non-coding RNAs: http://research.imb.uq.edu.au/rnadb
UCSC Genome Browser: http://www.genome.ucsc.edu
Access to this links box is available online.