

Tagging mammalian transcription complexity

Piero Carninci^{1,2}

¹ Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama 351-0198 Japan

² Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

The nature of the ‘transcriptome’ is more complex than first realized. Although CAGE, various tagging technologies and tiling arrays show that most of the mammalian genome is transcribed, a large proportion of transcripts do not encode proteins and are either poorly polyadenylated, involved in sense–antisense pairs or never leave the nucleus. In this article, I review the various techniques and data sets that are currently used to measure gene transcription and the evidence that reveals the true extent of transcription in mammalian genomes. The next few years will see efforts to identify novel transcripts systematically and decipher their function. A deeper understanding of transcriptional complexity might even lead us to redefine what we mean by the term ‘gene’.

Introduction

Now that the sequences of several genomes are complete [1,2], the focus of our attention has turned to the identification of the encoded genes and their function. Gene identification has relied on a combination of *in silico* predictions for evidence of transcription, such as ESTs, full-length cDNAs and already known genes. Conservative estimates suggest that there are ~20 000 protein-coding genes in the mammalian genome. However, there is growing evidence that RNAs, which are expressed at low levels and have less protein-coding potential, are also transcribed.

The development of whole-genome tiling arrays [3], cap-analysis gene expression (CAGE) and similar tagging technologies (Box 1) provide a new perspective on the number of transcripts and their variants. Early studies on human transcripts using tiling arrays with spotted PCR products covering human chromosome 22 [4] and oligonucleotide tiling arrays covering chromosomes 20 and 22 [5] concluded that, even when using stringent thresholds, half of the RNAs detected were novel (Table 1). Further systematic data sets have been produced with oligonucleotide tiling arrays to detect transcripts in several cell lines from human chromosomes 21 and 22 [6,7] and also high-resolution tiling chips covering unique regions of ten human chromosomes [8,9] (Table 1). In addition, sets of whole-genome tiling chips that cover the whole non-repeated part of the human genome with 36-nt probes [10] were used to detect transcription in the human liver. The mouse transcriptome, for which tiling array data sets are not yet available, has been

analyzed by sequencing full-length mRNA-derived cDNAs or CAGE, gene identification signatures (GIS) and gene signature cloning technology (GSC; see Glossary; Box 1; Table 1 [11]), enabling a similar conclusion to be reached: that most of the genome is transcribed.

Tiling arrays have been used to screen numerous loci in many different conditions comprehensively (Table 1). They can determine the internal structure of transcripts and therefore define the presence of introns and exons, whereas tag sequencing data provide greater resolution of the borders of transcripts, different transcriptional start sites (TSSs) and transcriptional termination sites (TTS), which cannot be unambiguously provided by tiling arrays (Figure 1). Although tag technologies have been used mainly for the mouse transcriptome and tiling arrays have been used mainly for the human transcriptome, these technologies are starting to reveal common messages.

Most of the mammalian genome is transcribed

Tiling and tag sequencing data agree that the genome is transcribed more than was previously thought. Tiling arrays of human chromosome 21 and 22 have suggested that the number of detectable transcribed exons that are expressed in at least one out of 11 cell lines is tenfold greater than the number of exons that are currently annotated (Table 1): although 26.5% of the tiling probes were positive in at least one cell line, only 2.6% of the probes are located within well-annotated protein-coding genes. However, most of the novel transcripts were detected in only one out of 11 cell lines, suggesting that transcription is frequently restricted depending on the cell type or the condition [7]. Tiling arrays also extend the exons of known genes: ~49% of the transfrag regions, which partially overlap a characterized exon, mRNA or EST, seem to cover genomic regions that are not covered by known exons and transcript borders. However, because the transfrags in these studies, obtained by labeling double-stranded cDNAs, do not identify the direction of transcription and can also include signal derived from antisense RNAs [7], experimental verification is necessary to clarify the transcript and exon borders. The whole-genome 36-nt oligonucleotide arrays cover all non-repeat elements of human chromosomes tiled on both strands (Table 1), enabling, together with a different probe labeling design, the transcript orientation to be identified. Even by adopting stringent bioinformatic filtering conditions to remove false

Glossary

Tiling arrays: microarrays designed to cover at regular interval whole chromosomes or genome regardless the genome annotation.

Transfrags: transcribed fragments identified by grouping the adjacent positively hybridizing probes of tiling microarrays, which are calculated based on the 'collective behavior of neighboring probes' method (for more information, see Ref. [7]). Owing to the high density of transcription that derives from both DNA strands, the complete structure of transcripts could be defined for transfrags mapping to known genes but not for those corresponding to novel genes.

TARs: novel transcriptional active regions identified by the Yale group, for which there is no other gene annotation. TARs are defined as series of at least five consecutive probes of 36 nt in the genome with fluorescence intensities that rank above the 90 percentile over all probes on the array [10] in a 250-nt window. TARs are used for novel uncharacterized transcribed regions.

Reference sequences (RefSeqs): curated mRNA sequences that can integrate multiple forms of mRNAs into most likely transcript models. This is an excellent resource for gene annotation but presents often a larger version of an mRNA, because it contains more extensive 5' and 3' ends than other GenBank entries and the literature [13].

TUs: transcriptional units group all of the full-length cDNAs and ESTs that have at least one overlapping base in an exon in the same orientation.

TK: a transcriptional framework that is derived from TU but must also share at least one splicing site, a TSS or a TTS.

Chromatin immunoprecipitation: ChIP is performed with antibodies that recognize DNA-binding proteins, for example, transcription factors, regulators and structural proteins (histones and their modified forms). ChIP can be followed by DNA chip experiments (ChIP-on-CHIP), quantitative PCR or tagging technologies.

positives, Bertone and colleagues found 14 884 genes in liver that were predicted *in silico* by Genscan but that did not have EST support were expressed; they also detected 10 595 completely novel transcriptional active regions (TARs) [10]. Only 41% of these TARs overlap with the chromosomes 21 and 22 transfrags [6] found using another platform either because the two platforms use different thresholds (at least five positive probes are required to define a TAR, but transfrags can be narrower), have different sensitivity and probe coverage or have a different false positive ratio.

Although previous experiments with tiling arrays had been performed on whole-cell fractions or cytoplasmic polyA+ RNA fractions, by using high-density chromosome 10 tiling arrays, Cheng and colleagues were the first to combine the use of high density arrays and polyA fractions. They used the polyA+ and polyA- fractions of both cytoplasmic and nuclear RNA from HepG2 cells, revealing a new transcriptional world [8]. They found that 41.7% of all the RNA transcripts were confined to the nucleus. In addition, in cytoplasmic-RNAs, less than a third are polyadenylated and so would have been missed by procedures that use the presence of polyA tails for transcript purification. The total number of detectable RNAs, including novel nuclear and cytoplasmic transcripts that are sufficiently stable to allow tiling chip detection (15.4%), exceeds the whole fraction of annotated protein-coding exons by one order of magnitude.

It is estimated that at least 62.5% of the mouse genome is transcribed as primary heterogeneous nuclear RNA from one or both strands of the genome in at least one of the many tissues analyzed [11]. This estimation was reached by measuring the length of genomic sequences between 5' TSS and 3' TTS identified by GIS, GSC, full-length cDNAs and pairs of 5'-3' ESTs sequences derived from the same

Box 1. Recently developed tagging technologies that target cDNA ends

Tagging technologies [62] derived from long SAGE [63] depend on producing short tag sequences close to 3' ends of the transcript, ligating the tags into concatamers, then cloning and sequencing the concatenated product. The location of these long SAGE tags can be identified in the genome sequence.

CAGE [64,65] requires the synthesis of cDNA from isolated RNA using either an oligo-dT or a random primer at high temperature (55–60 °C) in presence of trehalose and sorbitol, which confer thermal stability to the reverse transcriptase. After the first strand cDNA synthesis, the cap site on the 5' end of full length mRNAs is biotinylated and subsequently RNaseI is used to digest any single-strand RNA, including the biotinylated cap from non-full-length cDNA-mRNA partial hybrids. The biotinylated cap remains on the mRNAs-full-length cDNAs hybrids only. After selection with streptavidin magnetic beads followed by RNA hydrolysis, a linker containing the class IIS restriction enzyme *MmeI* is ligated to the 5' ends of first strand cDNA and used to prime second strand cDNA. *MmeI* recognizes the linker, but cleaves 20–21 bp inside the 5' ends of cDNAs. After ligation of a second linker to the 3'-end and various PCR and purification steps, these 20–21-bp fragments are concatenated with DNA ligase and cloned into a plasmid vector for large-scale sequencing. Depending on concatenation efficiency, up to 15 of the 20-bp long CAGE tags per clone are ligated, but usually 50 000–1 000 000 tags are sequenced [64,55,65]. The CAGE tags sequences are mapped to a unique location using a BLAST search against sequences in the databases (with ~60% efficiency). Ambiguously mapped tags map to more than one location (mostly to two to three), and a large proportion of these tags correspond to transcribed repeats. The enrichment over non-capped molecules (and so not full length or fragments) has been calculated to be some 330-fold. Alternative methods have been proposed [62].

GIS and GSC are two similar technologies that detect the transcript borders with 5'-3' ditag pairs by taking advantage of full-length cDNA selection as does CAGE, with the exception that only oligo-dT primers adapters can be used to ensure the presence of both 5' and 3' ends of the transcripts [66]. The oligo-dT primer contains the sequence of *GsuI*, another class-II S restriction enzyme just 5' of the oligo-dT sequence. Following the second strand synthesis, *GsuI* cleaves 16-bp outside its recognition sequence, thus removing the polyA sequence, which is then replaced with a linker containing a second *MmeI* site. The cDNA, containing *MmeI* sites at both ends, is then cloned into a plasmid (using the GIS procedure) or a modified λ (full-length cDNA) FLC vector that does not have size bias after subtraction or normalization, followed by bulk excision from λ phages into plasmids [67]. *MmeI* cleavage allows the removal of the central part of the cDNA, leaving only the 5'-3' ditags (corresponding to the 5' and 3' ends of the mRNA) bound to the plasmid backbone. After religation, these 5'-3' ditags are excised, concatenated and cloned followed by large-scale sequencing, similar to SAGE. Mapping the obtained ditags takes advantage of the presence of a longer cumulative sequence and the presence of tags that originate from both ends of the transcript.

oligo-dT primed full-length cDNA (Table 1). This corresponds to the primary transcripts including their introns, which are filtered out in the whole-genome tiling arrays because a large fraction of heterogeneous nuclear RNA fragments are readily degraded. Consistent with data from tiling arrays [6], transcription is often restricted to specific tissue or cell types: out of the 250 subtracted cDNA libraries from different tissues, ~96 000 cDNAs (53% of the 3' end clusters) were present in only one library [12]. Although cDNA collections, which derive mainly from total RNA preparation (including mainly cytoplasmic RNA), contain traces of nuclear RNAs [12], this fraction and the complete polyA- RNAs are clearly underrepresented

Table 1. Selected transcriptome data sets containing tags and cDNAs

Transcriptome data sets	Key findings	Tissue or cell	Refs
Mouse CAGE, 7.1 million other full-length cDNAs ^a	236 000 TSSs	>400 libraries from different tissues or stages, polyA- and capped RNA	[11]
Human CAGE 5.1 million, mostly from random primed cDNA ^a	180 000 TSSs	40 CAGE libraries various organs or lines, polyA- and poly+ RNA (mostly random primed)	[11,19]
cDNA, 158 000 with full sequence ^a GIS, >118 000 ditags ^a GSC, >968 000 ditags ^a 5'-3' cDNA ESTs, 722 000 pairs ^a	181 000 distinct mRNAs, 44 000 transcriptional units, 47% of which have protein-coding potential; 62.5% of the genome is transcribed	250 full-length cDNA libraries and six ditag libraries	[11]
Mouse 8.55 million LongSAGE tags Tiling arrays^b	24 000 novel loci	72 libraries	[18]
25-nt probes per 35-bp-spacing tiling arrays containing human chromosomes 21-22 (Affymetrix)	26.5% of probes are positive in at least one line	11 human cell lines, polyA+ RNA	[6]
As above	Sense and antisense RNA co-expression; transcription-factor-binding site	NCCIT cell line +/- induction with retinoic acid	[21]
25-nt probes in 5-bp-spacing tiling arrays containing ten human chromosomes (Affymetrix)	10.1% of probes are positive in at least one cell line (4.9% are common to all cell lines)	Eight cell lines, cytoplasmic polyA+	[8]
As above	15% of probes are positive; 41.5% of the transcripts are found in the nucleus	HepG2 liver cells, both nuclear polyA+/- and cytoplasmic polyA+/-	[8]
36-nt probes in 46-bp-spacing tiling arrays containing all human chromosomes (Yale)	10 595 novel TARs were not previously annotated; 14 884 transcripts were predicted by Genscan without other cDNA or EST support	Human polyA+ pooled liver RNA	[10]
300-nt PCR products containing human chromosome 22 (Yale)	53% of the positive probes are from unannotated regions	Placenta polyA+ RNA	[4]
60-nt probes in 30-bp steps containing human chromosomes 20 and 22 (Rosetta)	At stringent conditions, 2% of the probes are positive; 47% of them are found in introns or intergenic regions	Six and eight mRNA from cells and tissues	[5]

^aThe number of tags and full-length cDNA clones does not indicate unique transcripts but refers to the tags and clones in a large data set with various degrees of redundancy.

^bOligonucleotide tiling arrays cover the non-repetitive regions of the chromosomes.

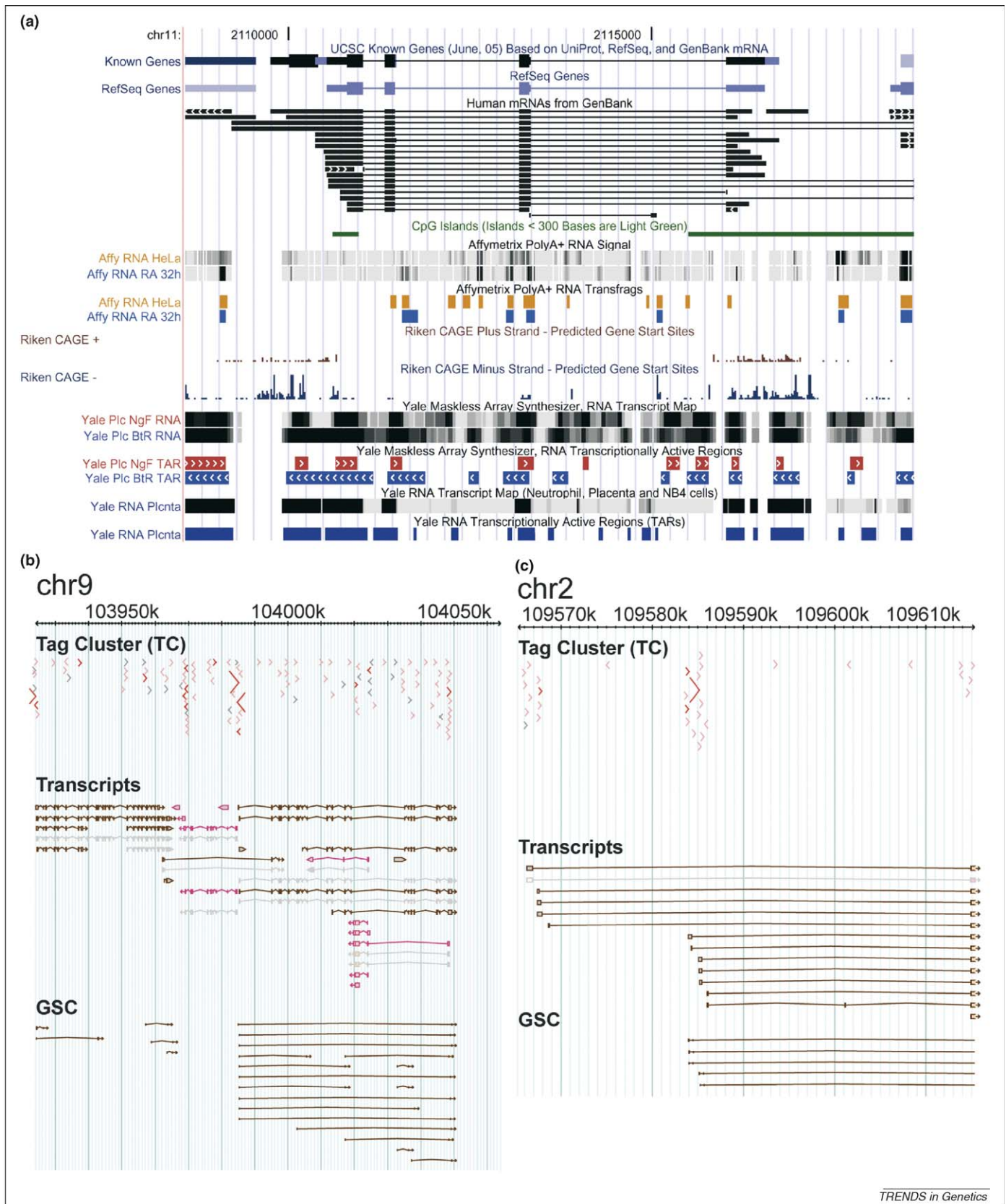
because full-length cDNA libraries from nuclear RNA fractions and polyA- fractions were never attempted, making a complete comparison not yet feasible.

Cross-validation of low frequency transcripts

Both the human tiling arrays and mouse tag methods identified an unexpected number of unannotated transcripts (Table 1 and Box 2). Tiling array signals from neighboring positive probes identified sequences of expressed RNAs that are part of the same exon. However, they cannot be unambiguously merged together to produce an accurate transcript structure nor can their ends be determined unambiguously. An initial tiling-tag comparison was performed by sequencing 997 000 CAGE tags [11] from whole RNAs from the human cell line HepG2. Only 20% of the Hep-G2 tiling array transfrags [6] had 'hits' corresponding to CAGE tags, although this was expected because CAGE tags tend to overlap with the 5' ends of mRNAs preferentially and a large part of the transfrag represents primarily internal exons. There is a dramatic decrease in hit rate when these tags are compared with transfrags obtained with high-resolution 5-nt-spaced tiling arrays [8], which decreases to 2% for the nuclear polyA- and to 3.3% for polyA+ cytoplasmic novel transfrags (Figure 2a) (defined here as the transfrags that map at least 1-Kb away

from reference sequence (RefSeq) mRNAs [13]). The transfrags overlapping the RefSeq mRNAs have a greater hit rate by CAGE (9-15%, depending on the cell compartment). If RefSeq transfrags are grouped together, >66% of these groups have overlapping CAGE tag hit(s) within a 100-bp distance (Figure 2a). The hit rate extends to >75% when hits within 1000-bp are considered. The dramatic difference in coverage between RefSeq and the rest of the transfrags suggests that novel transfrags detect RNAs that are expressed at less than one copy per million, that there are false positive tiling transfrags or false negative tags hits or that some transcripts do not have the cap structure. It is also possible to miss potential 'hits' by using CAGE tags that do not map to unique genome sites (~20% of the tags) because they were not used in the analysis [11]. Separate experiments to validate the outcome of these different technologies all identified novel transcripts (Box 2).

CAGE tags can be grouped in clusters of overlapping tags, which identify putative transcript start sites. Mapping these CAGE tags against the Hep-G2 high-resolution tiling arrays transfrags [8] suggests a remarkable correspondence: 74% and 92% of the HepG2 CAGE tags map within 100-bp and 1000-bp of the transfrags, respectively; non-mapping tags are preferentially located outside their putative upstream regions, indicating that the direction of



TRENDS in Genetics

Figure 1. Examples of transcriptome complexity. **(a)** An image of selected traces from the UCSC viewer (<http://genome.ucsc.edu/cgi-bin/hgTracks>) of the human insulin-like growth factor II region (on chromosome 11), a region of the Encode project, the tags and tiling arrays of which are available. Known genes, RefSeq and part of the human RNAs are shown in the upper part of (a). Here, the direction of transcription is from right to left (i.e. the lower strand is transcribed). Transcription of mRNA seems largely driven by a large CpG island. In the middle, there are some representative signals from Affymetrix (affy) tiling arrays shown in black and white (HeLa and RA 32 hours are shown), with their extrapolated transfrags shown below. CAGE tags are displayed in both of their orientation above the Yale 36-nt tiling arrays, where the black and white bars indicate the intensity of the signal; the transcriptional active regions (TAR), with the orientation available for the TAR of placenta (Plc) NgF and Plc BtR biological samples are shown below. Tiling arrays and CAGE identify more transcription and TSS than the cDNA or gene models, and include evidence of bidirectional transcription. The CAGE tags that originate in the 3' end of the transcripts represent short RNAs that might have regulatory functions. **(b)** A sketch indicating the complexity of a region of

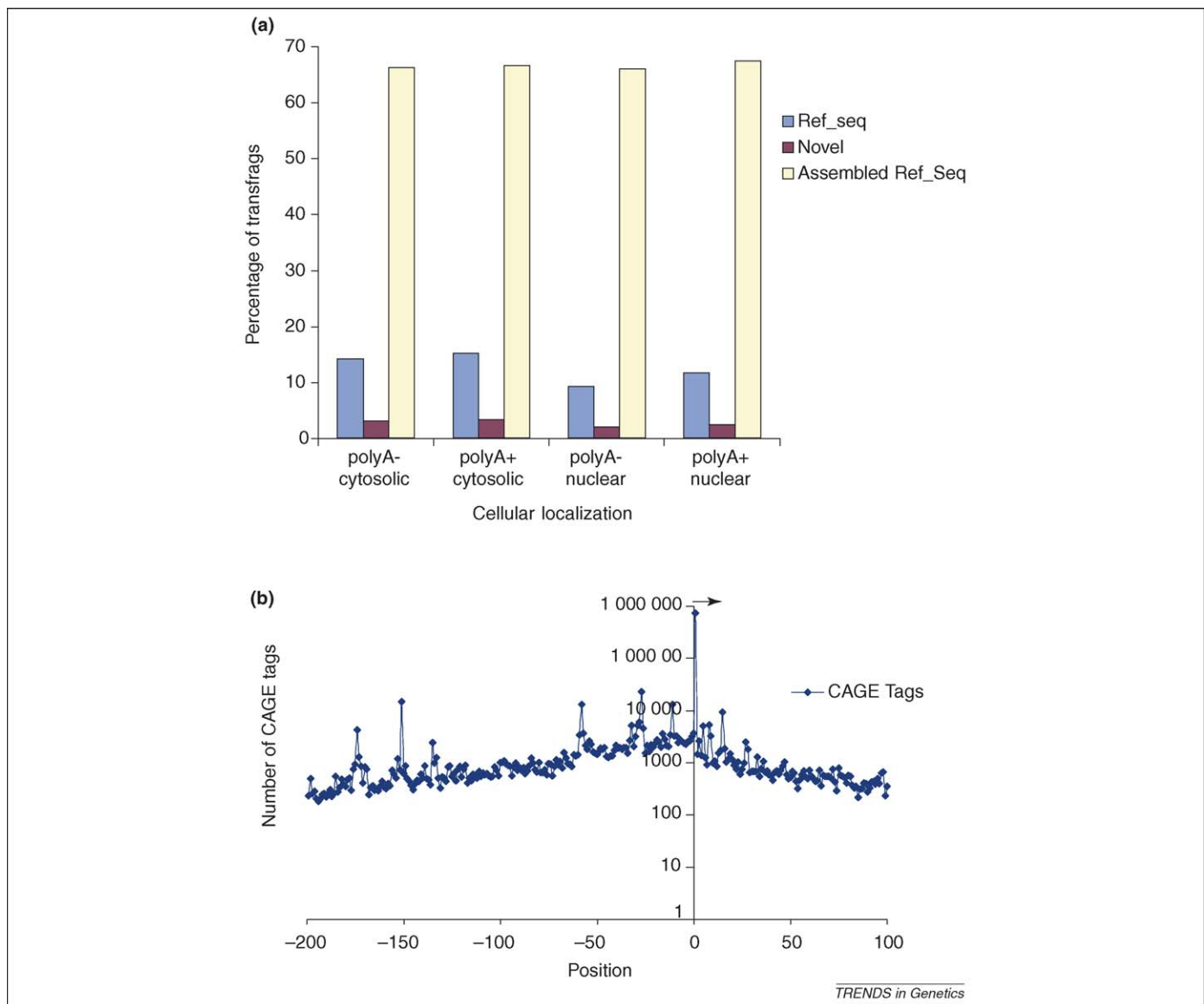


Figure 2. A comparison of the Hep-G2 CAGE data with 5-bp-spaced transfrags from ten human chromosomes [8]. **(a)** The number of transfrags from Hep-G2 with CAGE tags, represented separately for each cell compartment (i.e. cytoplasmic or nuclear), together with the polyA status (polyA+ or polyA-) is shown. Transfrags identified by RefSeqs are more frequently verified by CAGE tag. **(b)** Approximately 92% of the CAGE tags are localized within 1000 bp of all transfrags obtained (either nuclear and cytoplasmic polyA- or nuclear and cytoplasmic polyA+). The value 0, on the x-axis, indicates the overlap of CAGE with transfrags (an arrow indicates the direction of transcription of the CAGE tags); the values that are less than zero indicate the CAGE tags that identify transcription towards the transfrags; positive values indicate tags that map downstream of the closest transfrags and identify transcription departing from the transfrag regions. It is evident that there is preferential mapping of CAGE tags upstream of transfrags to define their TSSs.

transcription is towards the transfrags (Figure 2b). This suggests that clusters of tags reliably indicate the starting sites of mRNAs identified by transfrags, and help to refine the 5' end of these mRNAs (Box 2).

Many of these transcripts are present at low levels. Out of the 997 000 HepG2 CAGE tags, there are 97 353 distinct

TSSs that detect 24 423 transcriptional units. Among these TSSs, 65 501 map within 1-Kb upstream of known mRNAs or 5' ESTs, of which 41 974 are represented by a single CAGE tag, 7444 TSSs have only two tags hits and 16 083 contain three or more tags. These data indicate that most transcriptional events in a cell are rare: an average cell expresses

the mouse chromosome 9. Only non-redundant representative transcripts are shown. The transcripts and tags (GSC) that are transcribed from left to right are shown in brown, those transcribed from right to left are shown in pink, whereas the pale color indicates an Ensembl gene. The grey boxes indicate non-coding regions, whereas pale orange indicates coding regions (not visible for short exons). CAGE transcription is identified by the direction of the arrowhead. TSSs are indicated by color coded arrowheads: pink, one CAGE tag detected; gray, evidence is different from tags (5' ESTs, cDNAs); small red arrow, at least two CAGE tags or other evidence; intermediate and large red arrows indicate more than three and more than ten tags, respectively. GSC ditags indicate the beginning and the end of the transcripts only, suggesting the presence of multiple start-termination sites, some of which identify novel start-end pairs. In the samples analyzed, the transcripts were identified from left to right only (in b and c). Only representative GSC ditags are depicted. Notice the overlap of transcription, including an ncRNA bridging two otherwise independent genes. Because CAGE has greater coverage, it identifies several TSSs; there are 15 TSSs that have multiple evidence of transcription. Some of the reliable TSSs (red arrows) are in the exons and in the 3' UTRs. For more information, see <http://fantom.gsc.riken.jp>. **(c)** The *BDNF* gene, which has six different isoforms transcribed under the control of six different core promoters across a 22-Kb genomic region, five of which were identified by CAGE. Abbreviations: BtR, Bertone protocol; NgF, Nimblegen, forward; plcnta, placenta.

Box 2. The reliability of technologies

Lack of complete matching between tiling chips, CAGE and known mRNAs might be due to the nature of the data and experimental design; however, it is possible that these technologies detect technical or transcriptional background. To this end, independent RACE was performed to validate all of the technologies.

The predicted transfrags direction is not determined in tiling arrays; therefore, RACE was attempted in both directions by using randomly selected transfrag information for the primers design [8,9]. Transfrags were successfully validated in 82.6% of the cases and ~60% of the transcripts overlapped in both orientations. Using a high-confidence data set from human liver, the Yale group verified that 94% of the TARs identified by the 36-nt tiling arrays can be amplified. Using this data set can help explain the discrepancy with the previous studies [8,9].

In selected mouse genomic loci of specific biological interest where CAGE tags identified novel transcripts, RACE has confirmed the existence of transcription starting sites (TSS) in >90% of these transcripts [11]. To identify the TSS of *Oprm1*, RACE was performed for all of the tags, resulting in 33 different RACE clusters identifying different TSSs, RACE failed in five TSSs that were supported by one tag only. Although many of the single CAGE tags (singletons) are probably true TSSs, only the presence of two or more tags were considered reliable. RACE and RT-PCR validated each of the tagging and array technologies equally well, suggesting the existence of an expanded transcriptome.

some 400 000 mRNAs [14] and expression is not homogeneous among cells of the same type, but instead fluctuates [15]. The level of transcription that constitutes transcriptional leakage is still unknown. If these transcripts are biologically relevant, ideal platforms for their analysis would require detection of transcripts as rare as one in a million, which is achieved only in few cases, for example, Codelink and Agilent (<http://www.home.agilent.com>) [16], whereas the 25-nt Affymetrix GeneChip (<http://www.affymetrix.com>), the basis of the tiling chip, was reported to detect <55% of the rare mRNAs that are transcribed [17]. The sequencing of tag libraries is scalable and results show that up to 800 000 tags are required for statistical treatment of the expression of 70% of the transcription factors in the brain (Box 3; P. Carninci *et al.*, unpublished data).

Variability in transcript ends: are there true reference mRNAs?

5' end tagging analysis suggests the existence of 236 000 different TSSs and 153 000 TTSs in the mouse. Considering that only 44 000 transcriptional units (TUs) were detected in mouse transcriptome analysis, many TUs have considerable TSS and TTS variability [11]. Accordingly, long-serial analysis of gene expression (SAGE) in the mouse (Table 1) identified 3.3 alternative 3' ends per locus [18]. Because there is no complete overlap of tissues examined, the combination of starting-termination sites is likely to be much greater when all tissues, cell types and developmental stages are analyzed. The combination of transcripts that have multiple overlapping tags or evidence of TSSs and TTSs from full-length cDNAs suggest that there are at least 181 000 different transcripts, even without taking alternative splicing into account [11].

Alternative TSSs (Figure 1c) are important to detect core promoters that drive the expression of different mRNAs isoforms [19] in different contexts. By clustering tag-derived TSSs into 70 super groups based on their

Box 3. Technological and strategic implications

Different transcript isoforms are expressed at different stages in different tissue. Expression analysis that involves TSS usage by tagging technologies results in an efficient dissection of the promoter elements driving specific expression [19]. The ideal analysis would be a method capable of analyzing the whole transcriptome including the full-length cDNA structure and the expressions of each cDNA, which is not (yet) feasible. More realistic developments might take advantage of applying tagging methods [65,64,62] to a novel pyrosequencing-based instrument recently reported in a genome-wide TP53 mapping study [68] to identify rare transcripts and their TSSs in an comprehensive and affordable manner. Because sequencing costs are destined to decrease, tagging technologies become growingly appealing for large-scale expression analysis. Surely, sequencing and resequencing different transcriptomes that vary in different cell types is essential for a final understanding and annotation of different gene variants. Normalization and subtraction of cDNA libraries can still discover new transcripts [12] but these efforts should also be developed to include single cell comprehensive transcriptome analysis, because of the cellular complexity of whole tissues, and not limited to brain cDNA libraries.

Annotation of various genomes has been largely incomplete owing to the lack of extensive transcriptome data, under the assumption that conservation would be sufficient evidence for functional extrapolations. To add value to the current genome sequences, it is imperative to develop standard basic strategies to analyze the transcriptome for each sequenced genome comprehensively. Transcript identification will enable detection of the multitude of the regulatory elements, including the core promoters that are essential to deciphering true transcriptional networks.

The Encode project [69] is moving the field ahead by detecting not only transcript elements but also important functional elements by ChIP followed by tiling hybridization (ChIP on Chip) or coupled with tagging technologies. Important studies, going beyond the scope of this review, are beginning to provide essential information on regulatory elements such as transcription-factor-binding sites [70,24,68,21], regulatory proteins [71,72] and epigenetic modifications [73–75]. This and more information not cited here requires to be integrated to the current transcript and genes [48].

CAGE-determined level of expression, different TSSs belonging to the same TUs almost invariantly fall into different expression super groups that have different control mechanisms. This suggests that it is important to analyze the core promoter elements separately when studying transcriptional networks (Box 3).

CAGE data can lead to the identification of two types of promoters. Promoters with accurate TSSs are generally associated with TATA-boxes, are the minor proportion of the promoters and tend to transcribe tissue-specific transcripts [19]. However, for most of the promoters, RNA expression starts from broad regions that are usually associated with CpG islands and are depleted of TATA boxes. They have broad distribution of TSSs, usually spread over a 100-bp region, and the preferred start sites of these promoters are constituted by pyrimidine/purine dinucleotides, a simplified consensus of the initiator (Inr) element [19]. The longest transcripts in clusters, which are often used as RefSeqs [13], are not necessarily the representative transcripts and should not be the basis of core promoter element identification.

Chromatin immunoprecipitation (ChIP) analysis of the RNA polymerase II (Pol II) preinitiation complex [20] showed that >58% of hits detected by CAGE tag clusters were within a 100-bp region [19], which is a discrete

overlap considering that different starting samples were used. Incidentally, other ChIP-on-CHIP experiments suggest the existence of novel putative transcripts [20,21], which will require further experimental and bioinformatics integration (Box 3).

More complexity within known genes

There is good coverage of known RefSeq transcripts exons with both arrays and tags. Most RefSeq exons can be identified by tiling arrays and most TUs (>70%) can be identified by CAGE tags, despite largely incomplete overlap in the tissues examined with the two technologies [11]. However, CAGE tags detect both longer and shorter versions of annotated transcripts. Some of these transcripts partially overlap coding mRNAs [11], including the TSSs originating from within exons, which are preferentially associated with the coding exons of TATA-box promoted transcripts [19]. These RNAs might not encode proteins; however, transcription is conserved between mouse and human, suggesting that these seemingly shorter transcripts, or transcription itself, can be functional. Approximately 34 000 reliable TSSs, for which there was multiple evidence of transcription, were detected within coding sequences [19].

Although these RNAs could only be speculated as being involved in splicing regulation [22], it is noteworthy that ChIP mainly detects the hypophosphorylated form of RNA Pol II, which is associated with elongation and with alternatively spliced exons but not with the introns of human genes [23]. ChIP experiments and the complexity of TSSs (Box 3) show that transcription factors not only bind to putative promoter regions but also frequently bind to regions within or downstream of annotated genes [21,24].

Alternative splicing

Although tiling arrays identify several transfrags that are ten times larger than the exons of known genes and add exons to already annotated transcripts, the 'true' structure of all mRNAs and their isoforms is difficult to determine. The Affymetrix arrays do not identify 11% of the known exons because they are too short for the criteria used to identify exons. Eukaryotic exons also have frequent variations of the splicing site at the 3' end of exons by three, six or nine nucleotides (which would retain the reading frame but insert additional amino acids [25,26]) that are not easily detectable with tiling arrays. Defining the actual structure of mRNAs in the presence of multiple, alternatively spliced forms is a difficult task that requires either intensive full-length cDNA sequencing or the cloning of products obtained after rapid amplification of cDNA ends (RACE) [9]. For example, using alternative-splicing-enriched ESTs libraries to detect alternative splicing [27], 18 isoforms of the *Sorbs1* mRNA were identified in melanocytes and melanoma cells. From the Fantom-3 data set [11], Forrest *et al.* identified five alternative spliced forms of the *Csfl* receptor corresponding to membrane protein secreted forms and three alternative tethered forms, showing that splice variation alters the localization of the protein [28]. This occurs in >8% of coding TUs [11,29]. The mouse alternative splicing data suggest that

there are at least 78 000 different mammalian proteins [11] including splicing variants, which is probably an underestimation. (for a review on alternative splicing, see Ref. [30]).

Most novel transcripts do not encode proteins

There are only a few entirely novel protein-coding genes that have been discovered either by tiling arrays or tagging technologies. In the mouse, most novel TUs (7183) detected with full-length cDNAs in the Fantom-3 project [31] seem to be noncoding (Table 1) and, to date, the number of non-coding TUs (23 218) outnumbers the number of protein-coding TUs (20 929) [11]. Approximately 63% of the mouse non-coding RNAs (ncRNAs) are alternatively spliced [11], suggesting they are unlikely to be derived from genomic contamination during cloning procedures. However, their cross-species conservation is poor [32]; it is slightly greater than the background level of conservation between genomes and varies depending on the class of ncRNA [11]. This might suggest that at least a part of this novel transcription is transcriptional noise. However, well-characterized ncRNAs such as AIR and Xist have little conservation between mouse and human [33]. Conservation of long stretches of ncRNAs might not be required for their correct function because the conservation required need only be relatively short if they are to function as sense-antisense (A/AS) pairs [34,33].

Analysis of 2680 full-length ncRNAs showed that, although cross-species conservation of the transcribed region is weak, the conservation of their promoters is much greater and extends further than that seen in coding mRNAs (5 Kb versus 500 bp) [11]. In addition, the TSSs of both coding and ncRNAs are devoid of repeats and retrotransposon elements, which are also underrepresented in their immediate upstream putative promoter regions compared with those regions far upstream and downstream of the TSS [11]. Although long terminal repeat (LTR) elements can be used as promoter elements during specific developmental stages [35], the presence of repeats at TSSs seems to be selected against, possibly to avoid interference with chromatin-silencing mechanisms. These data suggest that these conserved upstream regions are the promoters of ncRNAs, and that they are not only expressed but also have dynamic variations in their expression as detected by RT-PCR [36], microarrays [37] and CAGE tags analysis [11].

The missing transcriptome

Even protein-coding RNAs might not be polyadenylated; studied conducted in the 1970s showed that hundreds of protein are produced from polyA- RNAs [38,39]. However, most cDNA libraries used for full-length cDNA projects have been prepared by oligo-dT priming on polyA+ mRNAs for technical reasons.

ncRNAs constitute another large proportion of the undetected non-polyA transcriptome. The Fantom-3 set of ncRNAs is transcribed by RNA Pol II; these ncRNAs appear in cap-selected libraries but are often poorly polyadenylated. Unlike protein-coding mRNAs that are cloned after polyA-based isolation on oligo-dTs, they are cloned through internal priming events >40% of the time. Large

ncRNAs, like AIR, are underrepresented in cDNA libraries and are mainly cloned as truncated cDNAs. Therefore, Furuno and colleagues searched for fragmented, non-coding cDNA clusters that are likely to be internally primed that map some distance from their promoters and therefore are located far from a CpG island. These cDNA clusters are likely to be fragments of unclonable large transcripts [40] that might be polyA- and therefore are poorly represented in other ESTs libraries. They have identified 66 putative long ncRNAs spanning 0.23% of the genome. Some were already known regulatory ncRNAs, such as UBE3A and Kcnq1, and 12 represented antisense (AS) transcripts. Experimental validation confirmed that 80% of them are true mega-transcripts with potential regulatory functions. Up to 2700 of such mega ncRNA might exist.

The new world of nuclear localized transcripts

Cheng *et al.*'s tiling array analysis emphasizes the extensive amount of nuclear transcription that occurs [8] and sheds some light on the old paradox that <5% of RNA synthesized by RNA Pol II is exported to the cytoplasm [14]. Approximately 25% of the nuclear restricted polyA-fraction was defined as intergenic relative to known protein-coding genes, whereas a large proportion of polyA- is enriched in intronic sequences (57%). Although the presence of intronic sequences can be thought of as trivial byproducts of splicing, intronic sequences can have additional functions and this could explain the slow degradation of some introns and the selective export of parts of them to the cytoplasm [8,41]. Intronic RNAs can also have regulatory function in the nucleus, such as splicing regulation [41]. In mouse, the imprinted, tandemly repeated small nucleolar RNA (snoRNA) HBII-52 is encoded in introns; however, in human it is encoded in exons, except for one case in which the snoRNA is encoded in an intron. It regulates alternative splicing of the serotonin receptor 2C, which is altered in Prader-Willi patients who do not express this snoRNA [42]. There are other well-known ncRNAs that never leave the nucleus, including Xist and Air, but there are no common functional mechanisms governing ncRNAs [41,43,44]. B2 RNA regulates the RNA Pol II during heat-shock [45]. The passage of the RNA Pol II machinery through certain genomic regions might be more important (e.g. in maintaining an open chromatin structure or displacing regulatory proteins from the chromatin) than the type of RNA produced. Transcription of *srg-1* RNA through the yeast *ser-3* promoter removes an activator protein and abolishes *ser-3* transcriptional activity [46]. Regulation in the opposite direction occurs in *Drosophila*, where transcription through a polycomb group response element is necessary to counteract silencing [47]; in humans, the failure to express transcripts from the FMR3 promoter, leading to mental retardation, can be caused by the lack of an ncRNA antisense to the FMR3 promoter [48]. Recently, S/AS transcription has been shown to regulate intergenic transcription by DICER and double-strand RNAs formation, for example, in the β -globin locus [49]. These examples highlight the variety and complexity of transcription-based mammalian control. However, their function is not restricted to the nucleus

and other examples indicate that there are more novel mechanisms. For example, NRON, a lowly expressed ncRNA represses the activity of NFAT, a transcription factor involved in T-cell-mediated immune response and development, by creating an RNA-protein complex that inhibits NFAT nuclear transportation [50].

Global sense-antisense transcription

Detection of S/AS has increased the coverage of the transcriptome. Although 15–20% of protein-coding genes were reported to undergo AS transcription [51–54], SAGE indicates that at least 50% of all transcripts have a corresponding antisense transcript [18]. However, we should be cautious about accepting this figure because reverse transcriptase, an enzyme used in the first-strand cDNA synthesis, has a tendency to produce spurious second-strand cDNAs, causing false positive antisense detection. This issue is addressed in various ways, including capturing the cDNA by their 5' ends by cap-trapping or oligo-capping [55], or strand-specific RT-PCR [34]. Related to this problem (and to the sensitivity issues), most of the Affymetrix transfrags data sets do not identify the orientation of transcription. However, when Kampa and colleagues employed a method that was less sensitive than that used by Affymetrix (by labelling the RNA directly), 11% of transfrags were found to have overlapping S/AS transcription [7]. Subsequent bidirectional RACE experimental validation of transfrags revealed that up to 60% were in an S/AS relationship, including S/AS pairs that seem to be produced by an unidentified putative RNA-dependent RNA polymerase [8]. By taking mouse CAGE tags that map in exons and introns into account, it seems that up to 72% of TUs are involved in S/AS transcription [34].

S/AS pairs are often differentially expressed across tissues and conditions, suggesting that they are regulated and unlikely to be transcriptional noise. Antisense ncRNAs expression level might be still underestimated. Although most mouse CAGE libraries were oligo-dT primed, a small set of random-primed CAGE libraries indicate an even larger antisense transcriptional level [34], consistent with observations that antisense transcripts are poorly polyadenylated [37].

SAGE analysis of human-mouse conserved S/AS pairs suggests that they seem to be co-expressed in the same tissue and that their expression is inversely regulated or discordant [56]. Other data sets do not give consistent results, because CAGE and RT-PCR validation of macrophage activation identify different complex patterns of coregulation [34].

Although various studies [53,57,58] suggest that the most common S/AS transcripts overlap in a tail-to-tail (or convergent) pattern, CAGE tags suggest that head-to-head (or divergent) overlap is more frequent because CAGE enables the detection of 5' ends, which are underrepresented in SAGE libraries. S/AS pairs are transcribed into the promoter of the other [34].

Further clues suggesting that S/AS pairs are functional derive from observations of different S/AS representation in different gene ontology (GO) categories [59]. For example, GO terms for 'intracellular localization' are significantly overrepresented as S/AS pairs, whereas the GO

terms for 'extracellular matrix' are underrepresented [34,57]. Divergent S/AS pairs are significantly overrepresented among transcription factors [34], suggesting that transcriptional interference is a possible control mechanism. S/AS transcription is almost universal in imprinted genes and in those controlled by candidate imprinted genes [60].

S/AS pairs negatively and positively regulate the expression of each other in 30% of the examples where experimental validation by perturbation has been performed [34]. The regulation is exercised either positively or negatively by one transcript on its counterpart. Some transcripts in S/AS pairs seem to regulate their counterpart, whereas others can only be regulated, showing the asymmetry of S/AS regulation. The existence of both negative and positive correlations of S/AS levels suggests that there are various mechanisms that are different from those in siRNA reciprocal control. Asymmetry of regulation might be influenced by yet another novel class of transcripts identified by CAGE as originating close to the ends of the 3' untranslated regions (UTRs), transcribed in the same direction as the full-length RNAs (Figure 1). These transcripts originate from a 'GGG' consensus sequence and are immediately followed at their 3' ends by conserved regions that are potential promoters [19]. When neighboring gene pairs that have increased 3' activity are located in a tail-to-tail pattern in the genome, their intergenic distance is significantly shorter (~2 Kb) than tail-to-tail transcripts that do not have 3' UTR transcriptional activity (~5 Kb). These ncRNAs might interact with the downstream AS transcripts, because GSC ditags have detected >1500 transcripts bridging the 3' UTRs of S/AS pairs [11].

What then is a gene?

The term 'gene' was originally coined to define chromosomal regions that influence a trait and was subsequently applied to the genomic elements that produce mRNAs, mostly associated with their produced proteins [61]. Although bacterial genes are amenable for such description, novel data show that our understanding of mammalian genes and genomes and their annotation are still overly simplistic. Novel data sets will ultimately affect the way we understand and study genes, because a large part of genome is transcribed from multiple positions and in both directions. The multitude of transcripts that exist have several TSSs, TTSs, alternative splicing variants and transcripts joining together regions that are traditionally annotated as different genes (Figure 1).

The genetics community has to rethink what is a mammalian gene in light of such profound differences with mammalian gene definitions in classic textbooks. Despite so much transcriptional complexity, it is imperative not only to complete the list of the transcribed parts, their coordinates, controlling elements and timing of expression but also to develop formal ways to treat each one of these. A first step to do this is to classify the overlapping sequences in TUs [11,51], which are further divided into transcriptional frameworks (TKs) if they share TSSs, TTSs or splicing sites. However, this is only a first step in categorizing related transcripts and further logical subdivisions

are required that take into account the presence and sequential order of transcriptional landmark elements for subsequent quantitative treatment of transcription. We are finally starting to grasp and measure the extent of the complexity of the transcriptome; only when we have done so will biology become a more measurable science.

Acknowledgements

I thank all of the members of the RIKEN GSC-GREG and GSL and Fantom-3 consortium members for data production, analysis, advice, discussions and support. I also thank A. Hasegawa, S. Katayama and S. Kondo for supporting transfrags and CAGE analysis; A. Sandelin M. Frith and C. Plessy for advice on this article; and Y. Hayashizaki for support and encouragement. This work was supported by a Research Grant for National Project on Protein Structural and Functional Analysis from MEXT, a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government and a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology.

References

- Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
- Johnson, J.M. *et al.* (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* 21, 93–102
- Rinn, J.L. *et al.* (2003) The transcriptional activity of human chromosome 22. *Genes Dev.* 17, 529–540
- Schadt, E.E. *et al.* (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* 5, R73
- Kapranov, P. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919
- Kampa, D. *et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14, 331–342
- Cheng, J. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154
- Kapranov, P. *et al.* (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* 15, 987–997
- Bertone, P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* 306, 2242–2246
- Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563
- Carninci, P. *et al.* (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* 13, 1273–1289
- Pruitt, K.D. *et al.* (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33, D501–D504
- Jackson, D.A. *et al.* (2000) The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *FASEB J.* 14, 242–254
- Ramsey, S. *et al.* (2006) Transcriptional noise and cellular heterogeneity in mammalian macrophages. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361, 495–506
- Hardiman, G. (2004) Microarray platforms—comparisons and contrasts. *Pharmacogenomics* 5, 487–502
- Draghici, S. *et al.* (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* 22, 101–109
- Siddiqui, A.S. *et al.* (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl. Acad. Sci. U. S. A.* 102, 18485–18490
- Carninci, P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635
- Kim, T.H. *et al.* (2005) A high-resolution map of active promoters in the human genome. *Nature* 436, 876–880

- 21 Cawley, S. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499–509
- 22 Bentley, D.L. (2005) Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.* 17, 251–256
- 23 Brodsky, A.S. *et al.* (2005) Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.* 6, R64
- 24 Loh, Y.H. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* 38, 431–440
- 25 Zavolan, M. *et al.* (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* 13, 1290–1300
- 26 Chern, T. *et al.* (2006) A simple physical model predicts small exon length variations. *Plos Genetics* 2, 606–613
- 27 Watahiki, A. *et al.* (2004) Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *Nat. Methods* 1, 233–239
- 28 Forrest, A.R. *et al.* (2006) Genome-wide review of transcriptional complexity in mouse protein kinases and phosphatases. *Genome Biol.* 7, R5
- 29 Davis, M.J. (2006) Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *PLoS Genet* 2, 554–563
- 30 Mendes Soares, L.M. and Valcarcel, J. (2006) The expanding transcriptome: the genome as the 'Book of Sand'. *EMBO J.* 25, 923–931
- 31 Carninci, P. and Hayashizaki, Y. (2006) Genome network and FANTOM3: assessing the complexity of the transcriptome. *Plos Genetics* 2, 492–497
- 32 Babak, T. *et al.* (2005) A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics* 6, 104
- 33 Pang, K.C. *et al.* (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22, 1–5
- 34 Katayama, S. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science* 309, 1564–1566
- 35 Peaston, A.E. *et al.* (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* 7, 597–606
- 36 Ravasi, T. *et al.* (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* 16, 11–19
- 37 Kiyosawa, H. *et al.* (2005) Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res.* 15, 463–474
- 38 Morrison, M.R. *et al.* (1979) Differences in the distribution of poly(A) size classes in individual messenger RNAs from neuroblastoma cells. *J. Biol. Chem.* 254, 7675–7683
- 39 Frith, M.C. *et al.* (2005) The amazing complexity of the human transcriptome. *Eur. J. Hum. Genet.* 13, 894–897
- 40 Furuno, M. *et al.* (2006) Clusters of internally-primed transcripts reveal novel long noncoding RNAs. *PLoS Genetics* 2, 537–553
- 41 Mattick, J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* 25, 930–939
- 42 Kishore, S. and Stamm, S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 311, 230–232
- 43 Mattick, J.S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet.* 5, 316–323
- 44 Mattick, J.S. (2005) The functional genomics of noncoding RNA. *Science* 309, 1527–1528
- 45 Wassarman, K.M. (2004) RNA regulators of transcription. *Nat. Struct. Mol. Biol.* 11, 803–804
- 46 Schmitt, S. and Paro, R. (2004) Gene regulation: a reason for reading nonsense. *Nature* 429, 510–511
- 47 Schmitt, S. *et al.* (2005) Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev.* 19, 697–708
- 48 Santos-Reboucas, C.B. *et al.* (2006) Lack of FMR3 expression in a male with non-syndromic mental retardation and a microdeletion immediately distal to FRA3E CCG repeat. *Neurosci. Lett.* 397, 245–248
- 49 Haussecker, D. and Proudfoot, N.J. (2005) Dicer-dependent turnover of intergenic transcripts from the human β -globin gene cluster. *Mol. Cell Biol.* 25, 9724–9733
- 50 Willingham, A.T. *et al.* (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* 309, 1570–1573
- 51 Okazaki, Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature* 420, 563–573
- 52 Kiyosawa, H. *et al.* (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* 13, 1324–1334
- 53 Yelin, R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* 21, 379–386
- 54 Werner, A. and Berdal, A. (2005) Natural antisense transcripts: sound or silence? *Physiol. Genomics* 23, 125–131
- 55 Harbers, M. and Carninci, P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods* 2, 495–502
- 56 Chen, J. *et al.* (2005) Genome-wide analysis of coordinate expression and evolution of human *cis*-encoded sense-antisense transcripts. *Trends Genet.* 21, 326–329
- 57 Chen, J. *et al.* (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.* 32, 4812–4820
- 58 Lehner, B. *et al.* (2002) Antisense transcripts in the human genome. *Trends Genet.* 18, 63–65
- 59 Harris, M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261
- 60 Nikaido, I. *et al.* (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res.* 13, 1402–1409
- 61 Snyder, M. and Gerstein, M. (2003) Genomics. Defining genes in the genomics era. *Science* 300, 258–260
- 62 Hashimoto, S. *et al.* (2004) 5'-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* 22, 1146–1149
- 63 Saha, S. *et al.* (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–512
- 64 Kodzius, R. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods* 3, 211–222
- 65 Shiraki, T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15776–15781
- 66 Ng, P. *et al.* (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2, 105–111
- 67 Carninci, P. *et al.* (2001) Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel λ -FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* 77, 79–90
- 68 Wei, C.L. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124, 207–219
- 69 ENCODE project consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640
- 70 Bertone, P. *et al.* (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res.* 13, 259–274
- 71 Lee, T.I. *et al.* (2006) Control of developmental regulators by polycomb in human embryonic stem cells. *Cell* 125, 301–313
- 72 Boyer, L.A. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441, 349–353
- 73 Bernstein, B.E. *et al.* (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169–181
- 74 Huebert, D.J. and Bernstein, B.E. (2005) Genomic views of chromatin. *Curr. Opin. Genet. Dev.* 15, 476–481
- 75 Roh, T.Y. *et al.* (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* 19, 542–552