The Journal of Experimental Biology 210, 1497-1506
Published by The Company of Biologists 2007
doi:10.1242/jeb.000406

Constructing the landscape of the mammalian transcriptome

Piero Carninci

Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan and Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan

e-mail: rgscerg@gsc.riken.jp

Accepted 16 January 2007

Summary

The principal route to understanding the biological significance of the genome sequence comes from discovery and characterization of that portion of the genome that is transcribed into RNA products. We now know that this 'transcriptome' is unexpectedly complex and its precise definition in any one species requires multiple technical approaches and an ability to work on a very large scale. A key step is the development of technologies able to capture snapshots of the complexity of the various kinds of RNA generated by the genome. As the human, mouse and other model genome sequencing projects approach completion, considerable effort has been focused on identifying and annotating the protein-coding genes as the principal output of the genome. In pursuing this aim, several key technologies have been developed to generate large numbers and highly diverse sets of full-length cDNAs and their variants. However, the search has identified another hidden transcriptional universe comprising a wide variety of non-protein coding RNA transcripts. Despite initial scepticism, various experiments and complementary technologies have demonstrated that these RNAs are dynamically transcribed and a subset of them can act as sense–antisense RNAs, which influence the transcriptional output of the genome. Recent experimental evidence suggests that the list of non-protein coding RNAs is still largely incomplete and that transcription is substantially more complex even than currently thought.

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1497/DC1

Key words: transcriptome, full-length cDNA, mRNA, non-coding RNA, genomics.

Influence of old assumptions

Sequencing the genomes of human and other so-called 'model' organisms paves the way for a holistic approach to understand biological phenomena, provided that a genome sequence is properly annotated for the genes it contains. In parallel with the feasibility studies leading up to the genome sequencing projects, a shortcut was developed to the core of the problem: identifying the *expressed* part of the genome, the messenger RNAs (mRNAs), by sequencing randomly picked complementary DNA (cDNA) clones. Early work focused on characterizing these clones with a single-pass sequencing run from one end of the molecule to give 'expressed sequence tags' (ESTs). These could be linked to the corresponding part of the genome sequence, thereby identifying the genes encoded within the genome. In fact, *ab initio* gene prediction algorithms (see Glossary) have proved to be particularly poor in identifying encoded genes and, despite later improvements, they are still based on the assumption that the relevant output of the genome consists of mRNAs.

Early estimates in the 1970s of the number of such cellular RNA species were 70 000 to 100 000, based on the kinetics of *in vitro* renaturation of mRNA/cDNA. All of these RNAs, whose number is substantially larger than the known protein-coding genes within the genome (~25 000), were thought to encode for proteins. Although the research community has long been aware of the existence of some hundreds of non-protein coding RNAs, these have been generally dismissed as exceptions to the widespread belief that non-structural RNAs would all be protein-coding, and recognized by the incorporation of polyadenosine (polyA) tags at the 3′ end of the mRNA molecule. Some mRNAs found to lack polyA tags were described in the late 1970s, but again were treated as exceptions and not pursued.

Clarifying the number of protein-coding genes, and the identification and meaning of non-protein coding RNAs, has required the development of novel technologies, starting with cloning methods that crucially incorporate the full length of the mRNA molecule, rather than a shorter, artefactual fragment.

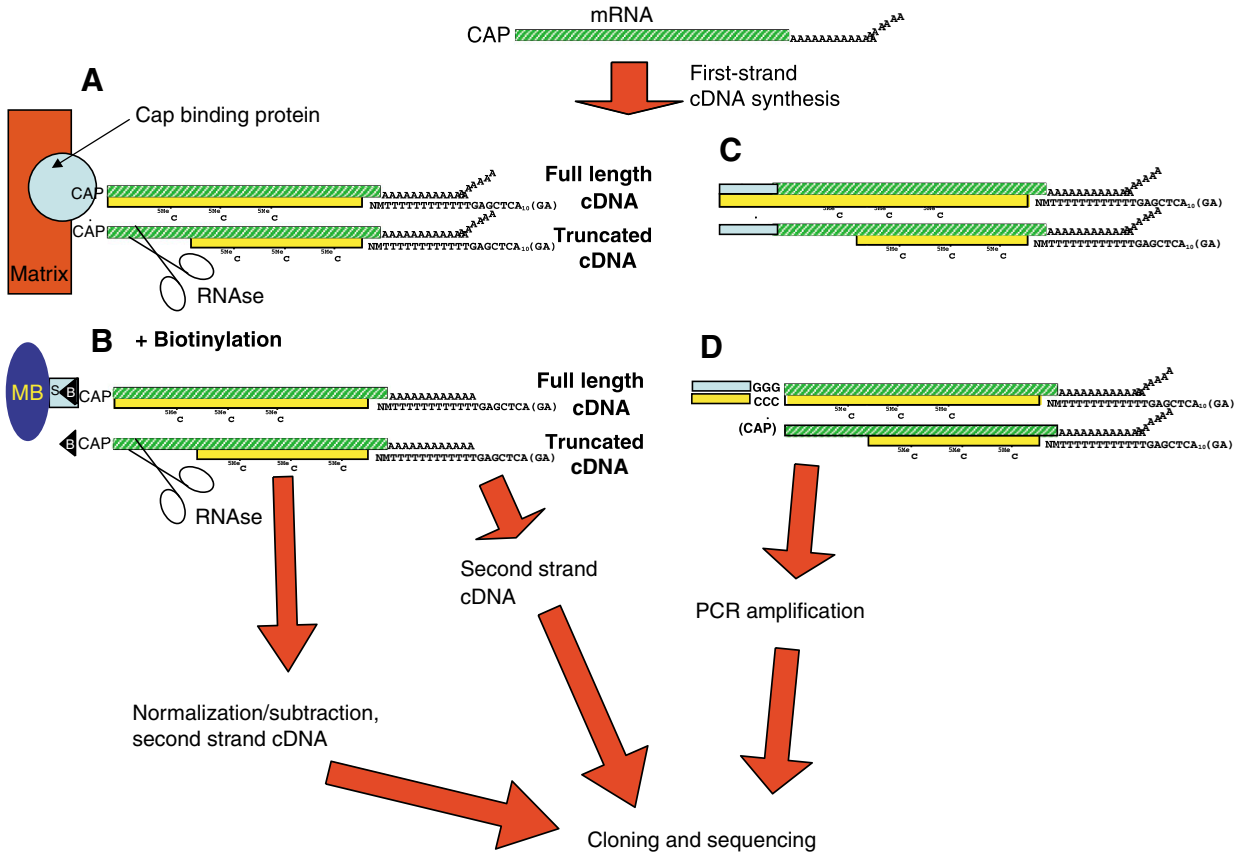

Fig. 1. Schematic representation of different methods for preparing full-length cDNA libraries. Starting from mRNA (top, in green with a polyA tail), first strand cDNA synthesis is associated with or followed by various options (A–D). (A) The resulting full-length cDNA/RNA hybrid (cDNA in yellow) is captured with a cap-binding protein; in the case of truncated cDNA/RNA hybrids the cap is removed by prior RNAse digestion. The cap-binding protein is immobilized on another support. (B) The cap-trapper protocol. A biotin molecule is added to the cap site; as in A, RNAse I removes the cap from the truncated cDNA/RNA hybrids, and the remaining full-length hybrids can be captured by streptavidin immobilized onto a support. After B, optionally, cDNA can be normalized and subtracted, otherwise after A or B it is denatured, subjected to second-strand cDNA synthesis and directly cloned. (C) The oligo-capping procedure, where an oligonucleotide is ligated to the mRNA instead of the cap structure. (D) The SMART oligonucleotide (a short, extended template at the 5′ end of the RNA template; see text for details) is also copied into the cDNA. Priming the oligonucleotide at the 5′ ends only allows for full-length cDNA selection. After C and D, PCR amplification is required before cloning and sequencing.

Despite an earlier report of a full-length cDNA library prepared by selecting full-length cDNA/mRNA hybrids through the modifications made on the 5′ end of the mRNA (Theissen et al., 1986), it was not until the middle of the 1990s that the need to systematically prepare full-length cDNA libraries was recognized. These libraries allowed the systematic discovery of the entire length of the coding mRNA, including its non-protein-coding ends (see Glossary). Conventional cDNA libraries, not enriched for full-length, have an average content of full-length cDNAs of 20–30% (Marra et al., 1999), while in high quality, full-length cDNA libraries, the proportion of full-length clones can exceed 90%. These libraries have thus become very attractive for large-scale sequencing projects, because they yield the full sequence data at a fraction of the sequencing cost of the entire genomic DNA, and because the greater consistency of the full-length sequence greatly aids data analysis, clone management and full-insert sequencing.

Here I discuss the methods used to produce full-length cDNAs in the RIKEN mouse project, and how this has generated a profound understanding of the scale and complexity of the transcribed genome (see also Glossary for the definitions of some elements of the transcriptome, which is the full set of mRNA molecules or transcripts produced). This project, along with others, has substantially contributed to a mature appreciation of even greater diversity in the transcribed genome that does not relate to protein coding, but probably to complex regulatory processes that underpin the generation of biological complexity.

Full-length cDNA cloning for gene discovery

Several full-length cDNA cloning approaches have been described to date, and are extensively reviewed elsewhere (Das et al., 2001); some of these have gained widespread usage (Fig. 1). The breakthrough in full-length cDNA production took advantage of a specific feature present at the 5′ end of

RNA molecules produced by the RNA polymerase II complex, namely the cap-site (Miura, 1981; Banerjee, 1980). This comprises an inverted (3′–5′) pppG nucleotide, which is added to the 5′ end of the polymerase II transcripts at very early stages of RNA synthesis.

The widely used SMART^R method of cDNA production is based on the addition by MMLV reverse transcriptase (RT), corresponding to the cap structure, of a trinucleotide CCC, which is annealed by an oligonucleotide having a GGG-3′ end. Use of the reverse transcriptase to synthesize on this cap-switch primer provides the means of priming the second strand of full-length cDNAs only (Zhu et al., 2001). Due to the relatively low efficiency, the polymerase chain reaction (PCR) is required. However, although these libraries are efficiently enriched for full-length cDNAs, they show a dramatically reduced variety of transcripts (less than half) when used for large-scale ESTs projects (Sasaki et al., 1998) if compared with non-PCR amplified full-length cDNA libraries prepared from the same tissue (Carninci et al., 2003). The oligo-capping procedure (Maruyama and Sugano, 1994; Kato et al., 1994) is more sophisticated. Uncapped RNA molecules, such as truncated RNAs, ribosomal and other structural RNAs, are dephosphorylated by a phosphatase. Next, the removal of the cap structure by tobacco acid pyrophosphatase leaves a phosphate group at the 5′ end of full-length mRNAs only, to which an oligonucleotide is added by RNA ligase, followed by library preparation by reverse transcription (RT) and PCR. Despite the requirement for PCR, this method has been widely used for the production of various cDNA collections including the full-length Japan (FLJ) human cDNA collection (Ota et al., 2004).

To clone a large variety of mRNAs efficiently without PCR, new full-length cDNA cloning approaches have been developed based on the separation of full-length cDNA from the artefactual truncated cDNAs by full-length cDNA/mRNA selection through the cap-site, while RNAse digestion cleaves the single-strand portion of the mRNAs, which happens when RNA is not protected by full-length cDNAs extending to the cap-site (see Fig. 1). RNAse removes the cap-site from these truncated cDNA/RNA hybrids. Full-length cDNA–RNA hybrids can then be physically selected using selection techniques based on retention of the cap structure. This can be achieved by direct binding of the cap with a cap-binding protein (Edery et al., 1995) (see Fig. 1A) which, however, requires tedious coupling of a mammalian cap-binding protein to a matrix and requires a substantial amount of starting mRNAs. Alternatively, the cap can be selected after its chemical modification by the addition of a biotin, followed by selection with streptavidin-coated magnetic beads (Carninci et al., 1996; Carninci et al., 1997; Carninci and Hayashizaki, 1999) (see Fig. 1B). This technology, called ‘cap-trapper’, makes use of commercially available reagents to oxidize the diol group at the cap site with $NaIO_4$, followed by biotinylation with a long-arm biotin hydrazide, which is very efficient and allows further manipulations downstream without using PCR, even if starting with as little as ~1.5 μg of total RNA (Carninci et al., 2003).

Comprehensive genome annotation requires unbiased cDNA cloning

Development of the full-length cDNA isolation technologies was only the first tool necessary. Although full-length libraries proved satisfactory in terms of full-length rate (~95%) (Carninci et al., 1996), they were not ideal for efficient isolation of difficult RNAs. In fact, the efficiency of conversion of mRNA to full-length cDNAs, and subsequent cloning, was inversely proportional to the length of the original RNAs, with clear under-representation of cDNA deriving from long mRNAs. This problem can be partially obviated by the use of engineered reverse transcriptases (RT), which have been altered by mutating the RNAseH domain (for instance, Superscript II and III from Invitrogen). Together with the use of these enzymes, we have found that some small molecules, also called osmolytes, which are synthesized by a multitude of organisms including yeast under conditions of stress (De Virgilio et al., 1994; Hottiger et al., 1994), effectively activate RTs at a high temperature (60°C) that would normally be inactivating. This enzyme ‘thermoactivation’ is promoted by the addition of trehalose and sorbitol to the reaction mixtures (Carninci et al., 1998; Carninci et al., 2002), enabling the preparation of cDNAs that exceed 15 kb in length.

Conventional plasmid vectors are strongly biased to clone short cDNAs present in cDNA ligation mixtures preferentially. This generates short insert libraries on average [(1–1.5 kilobases (kb)], even when the input molecules of cDNA are of a larger average size (>2.5 kb). To overcome this problem, cDNA mixtures containing such long cDNAs can be cloned into lambda vectors specifically designed for long cDNA cloning. These lambda FLC (full-length cDNA) vectors were derived by adjusting the size of the vector to just below the nominal cloning capacity (37.5 kb): the lambda phage most efficiently packages DNA of lengths close to the wild-type size (48.5 kb), so large cDNAs that traditionally were unclonable can now be packaged and cloned more efficiently than shorter cDNAs (Carninci et al., 2001). This has enabled the preparation of comprehensive cDNA libraries of size of 2.5–3 kb. Such libraries yield up to twofold greater diversity of cDNAs by random sequencing compared to libraries of shorter size.

Targeting rare RNAs

The ultimate tool available for maximizing gene discovery by sequencing of randomly selected cDNA clones is to remove undesired cDNA sequences through normalization and subtraction by hybridization (Bonaldo et al., 1996). In mammalian cells and tissues, the RNAs can be divided into classes of expression. Relatively few genes may account for up to 20–30% of the total mass of the mRNAs, whereas intermediately expressed (1000–2000 different RNAs) and rarely expressed (>10 000 different RNAs) gene classes account for the remaining 30–50% and 30–40% of the cellular RNAs, respectively. Although the proportions of these RNA classes vary in different tissues and cell types, in order to avoid prohibitive scaling up of sequencing operations, it is mandatory

to reduce the frequency of the highly and intermediately expressed RNAs and increase that of the rarely expressed sequences. Since the cap-trapper protocol is efficient, we developed methods to rebalance the frequency of transcripts representing different genes (normalization) and, secondly, to remove from the library those cDNAs already collected (subtraction). Indeed, use of cap-trapped, normalized/subtracted cDNA libraries is much more efficient for the discovery of novel cDNAs (Carninci et al., 2000; Hirozane-Kishikawa et al., 2003).

Subtraction and normalization have been widely used to produce diverse EST libraries rich in novel transcripts, and also for gene discovery in many organisms including human (Hillier et al., 1996; Marra et al., 1999) and rat (Scheetz et al., 2004). These libraries have contributed substantially to our current knowledge of gene structure and its many variations in mRNAs, and for full-length cDNA-based ESTs (Carninci et al., 2003).

Significantly, normalization and subtraction protocols tend to select against alternative splicing variants (different mRNAs generated from the same coding sequence by alternative selection of coding modules contained within it), and these have been discovered mainly by accident as hybridization leftovers. Although in the mouse transcriptome we have already identified more than 78 000 different splicing variants out of 44 000 transcriptional units (TUs; a TU groups together all of the mRNA sequences that show transcription overlap, see Glossary) (Carninci et al., 2005), splicing diversity is expected to be much larger. The comprehensive discovery of splicing variants necessitates different approaches, some of which may take advantage of selection of mis-paired nucleic acid hybrids (Watahiki et al., 2004; Thill et al., 2006). Besides displaying alternative exons, however, new methods will have to include full-length cDNA cloning, because it is not possible to reconstruct the structure of full-length mRNA transcripts without full-length cDNAs.

Coverage is far from complete

Subtracted/normalized full-length cDNA libraries have allowed extensive coverage of the transcriptome. While producing approximately 2 million ESTs, we monitored the subtraction rate during production of each library. We removed more than 90% of the abundant or already collected cDNAs in a large part of libraries, and then calculated that 13.9 million EST sequencing passes would have been required to achieve the same coverage and capture of rare RNAs using conventional libraries (Carninci et al., 2003), thus representing a considerable saving in time and money. Although many consider the coverage of the mouse transcriptome is close to saturation, it is important to note that the continued introduction of new tissues and stimulated cell types has provided a continuing high novel gene discovery rate that shows no sign of levelling out (Fig. 2) (Carninci et al., 2003). For instance, sequencing the 3′ ends of 15 000 macrophage cDNAs from a subtracted library still yielded >20% new clusters. This is remarkably high, considering that sequences were achieved in the late stage of a large-scale project (after producing ~90% of the RIKEN ESTs), when one would assume that most genes should have been already discovered (Carninci et al., 2003). Although we have now discovered a very large number of transcripts [>181 000 (Carninci et al., 2005)], which exceeds even the largest estimate (120 000) of the number of genes (Liang et al., 2000), and we have shown that 62% of the genome is transcribed into primary RNA transcripts, we have still not yet isolated all the RNAs that could be discovered using this approach (Carninci et al., 2003). Similarly, the rat EST project shows that the identification of novel genes using subtracted libraries was still yielding a considerable number of novel cDNAs at the moment of its publication (Scheetz et al., 2004).

Tiling arrays identify large RNAs complexity

To assess the complexity of the transcriptome without cDNA cloning, whole-genome tiling arrays have been developed. These provide an evenly distributed series of oligonucleotide array probes designed from the genomic regions not covered by repeat elements (reviewed in Mockler et al., 2005; Carninci, 2006). The mRNAs (or non-cloned cDNAs) isolated from tissues are labeled and hybridized to these arrays and the expressed regions of the genome identified from the distribution of positive array probes. These expressed regions are bioinformatically grouped into contiguous expressed regions, which are either called ‘transfrags’ (transcribed fragments) in the Affymetrix platform (Cheng et al., 2005) or TAR (transcriptional active regions) with the Yale platform (Bertone et al., 2004). Regardless of platform differences (Mockler et al., 2005), human whole genome tiling has demonstrated that a large part of the genome is transcribed into stable RNAs (~25%) and that a large part of the transcript is cell-specific, as almost half of the novel transcripts (and 20% of the known transcripts) are specific for only one cell line out of eleven tested (Kampa et al., 2004). This is in agreement with the full-length derived ESTs (Carninci et al., 2003), suggesting that the number of identified transcripts will rise simply by increasing the number of tissues and cells analyzed, although it is hard to define the plateau using current data. In particular, isolation of RNAs from minor cell populations within large tissues has not yet been properly addressed.

Even more surprisingly, the number of mRNAs that lack poly-adenylation is as large as the number of polyadenylated RNAs (Cheng et al., 2005) and more than 41.5% of the RNAs are confined to the nuclear regions. As such RNAs were never considered for gene discovery use and there are no *ad-hoc* technologies for cloning them, we can assume that transcriptome complexity is at least some fourfold larger than our current description based upon full-length cDNAs and ESTs (Table 1), which were derived from polyA-plus RNA isolated from whole RNA enriched for cytosolic RNAs.

CAGE tags suggests large number of transcripts and their variants

Cap-analysis Gene Expression (CAGE) technology uses the cap-trapping as the first step to capture the 5′ ends of the

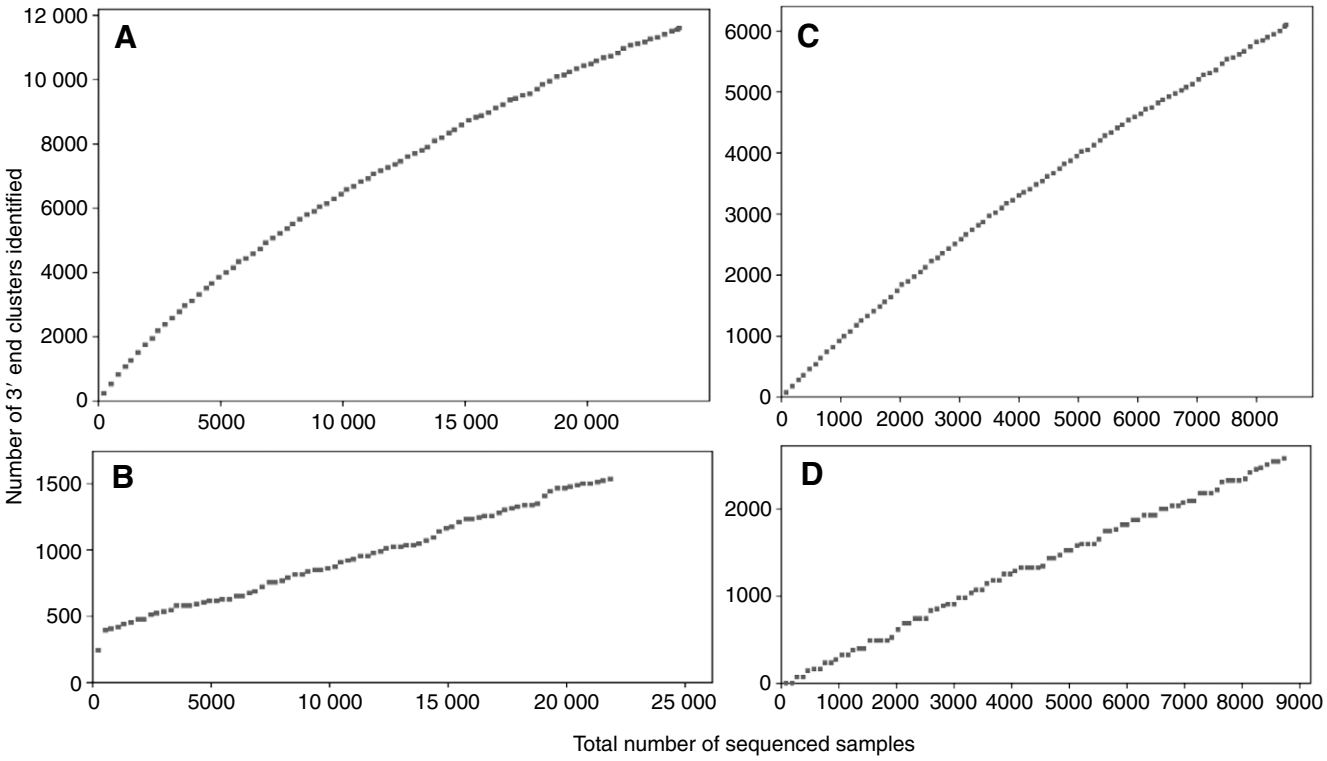

Fig. 2. Transcriptome discovery does not reach a plateau. (A–D) Rate of gene discovery as a function of the accumulated number of sequences. Gene discovery is represented from the number of 3′ end clusters identified. A and B refer to a mixed cell line library; C and D to an embryo pituitary subtracted library (Carninci et al., 2003). (A,C) Internal redundancy in terms of novelty of 3′ ESTs (*y* axis) compared to the sequenced cDNAs (*x* axis) within the library. (B,D) Novelty rate of these two libraries *versus* the entire RIKEN database of 1.4 million 3′ ESTs (axes as in A and B). Note that the discovery rate, although different, is far from reaching a plateau, suggesting the existence of a much larger number of yet undetected RNAs in each tissue/cell line. The pituitary gland library was prepared at the end of the project, and so far has identified more than 2000 new transcripts.

cDNAs, which are then transformed in short sequence (tags) of 20 nucleotides (nt) corresponding to the mRNA transcriptional starting sites (TSS) (Kodzius et al., 2006; Shiraki et al., 2003; Harbers and Carninci, 2005). We have produced millions of mouse and human CAGE tags (Carninci et al., 2006). Unpublished CAGE analyses suggest that in the human HepG2 cell line, used to produce close to one million of CAGE tags, there are about 66 700 TSSs mapping close to the first exons of known TUs, which can therefore be considered true 5′ end candidates deriving from full-length mRNA transcripts. Of them, about 47 000 appeared only once, while only 7700 were represented by two tags, and 12 000 by three or more tags. This suggests that the majority of the different, rarely expressed transcripts require an analytical technology with enough throughput and sensitivity to detect at least a million different transcripts for each cell type, a number tenfold larger than the current sequencing capacity (Carninci et al., 2005) and larger than current estimates of transcriptome diversity (Jackson et al., 2000). Notably, data derived from analysis of CAGE tags are largely confirmed by whole-genome tiling arrays (Carninci, 2006), while the opposite is not always true. This may possibly be due either to false positive signals with tiling arrays, or because the transfrags may identify uncapped RNAs that are not detected by cap-selection based methods.

Tagging technologies (Harbers and Carninci, 2005) have been developed with a sensitivity at least one order of magnitude larger than EST sequencing to detect transcripts, exhaustively to identify transcripts (Ng et al., 2005), identify their promoters, and correlate them with expression profiling by counting the tags as a digital measure of gene expression (Harbers and Carninci, 2005; Nilsson et al., 2006). Unexpectedly, these technologies have also revealed a surprisingly large degree of fine variability of transcription start and termination sites (Carninci et al., 2005). In the mouse, we have grouped all the transcripts in 44 000 transcriptional units (of which less than 21 000 are protein coding). By taking the conservative approach of requiring independent evidence for both the TSS and TTS (transcription termination sites) *via* analysis of their starting and termination sites, more than 181 000 independent transcripts were identified in mouse, whereas there are at least 238 000 independent TSSs and 153 000 TTSs.

This variability in TSSs highlights biologically significant differences between TSSs contained within a single TU, and indicates enormous complexity in the mechanisms mediating their regulated expression. For instance, CAGE analysis has identified promoters in the 3′ UTRs of many genes (Carninci et al., 2005). When two genes map tail-to-tail on the genome

Table 1. *Potential number of different RNAs in mammalian*

Variability type	Minimal estimation	Projection	Reference
Transcription starting sites (TSS)	236 000	>500 000	(Carninci et al., 2005; Carninci et al., 2006)
Transcription termination sites (TTS)	153 000	>180 000	(Carninci et al., 2003)
Tissue specific mRNAs	~half the transcripts are cell specific (11 lines)	Unknown for all the cells	(Kampa et al., 2004)
Large, non-polyA RNA	(Not possible to unambiguously group into individual RNAs)	Double the number of the RNAs above	(Cheng et al., 2005)
Nuclear specific	(Not possible to unambiguously group into individual RNAs)	Double the number of the RNA above	(Cheng et al., 2005)
Short RNA (miRNAs class)	>3000	20 000 (mouse) 70 000 (*Arabidopsis*)	(Mineno et al., 2005) (Lu et al., 2005)
Short RNA (but longer than 25 nt)	>100 clusters	Thousands (testis)	(Kim, 2006)
Splicing difference	Including splicing (78 000), more than a million	Not available	(Carninci et al., 2005)

Final estimation of the transcript number is not possible, but may be derived by combining the data obtained from the various modalities of RNA expression (TSS, TTS, tissue specificity, polyA status, compartmentalization, size and splicing).

(i.e. the 3′ ends of genes mapping in opposing genomic strands are terminating towards each other), the rate of 3′ UTR transcription is higher when two genes map closer to each other (average gap of ~2 kbp) than for tail-to-tail genes having low 3′ UTR transcription (~5 kb). Other genes, which do not map as tail-to-tail, also show 3′ UTR transcription, but no clear patterns are evident. In all cases, such 3′ UTR transcripts have true, conserved promoters that can activate transcription of a reporter gene (Carninci et al., 2006).

CAGE tags allow the classification of the TSS clusters into two main categories, based on the shape of the TSS. Surprisingly, the largest category of mammalian promoters does not show an accurate TSS, but instead a broad TSS (spread on average over up to 100 bp), generally associated with promoters constituted by CpG islands (see Glossary). Within such CpG islands, transcription starts mostly from pyrimidine/purine dinucleotides, a simplified consensus of the ‘initiator’ element, and these promoters are generally devoid of TATA-boxes (see Glossary). A much smaller fraction of promoters show well-defined, sharp peak TSSs, which are located 29–32 nt downstream of a classic TATA-box. Genes having TATA-box promoters are also preferentially associated with the presence of unusual transcripts, originating from exons (Carninci et al., 2006) (reviewed by Sandelin et al., in press). These exonic transcripts might consist of non-protein-coding regulatory RNAs, which are speculated to influence the chromatin status. Except for the brain, TATA-box promoted transcripts tend to be tissue-specific (Gustincich et al., 2006), whereas CpG, broad promoters seem to be involved in tissue-specific transcription, suggesting in turn that epigenetics features are particularly relevant for brain transcriptional control. Elsewhere, CpG promoters generally promote the transcription of housekeeping genes. The promoter shape can be defined only when many CAGE tags are identified (>100 per cluster), which happens in cases of highly and broadly expressed transcripts (8100 mouse and 6900 human promoters); however, all datasets described above have pointed at the existence of RNAs that are rare and specifically expressed, for which such general promoter properties analyses will require larger CAGE datasets.

Full-length cDNAs have been instrumental in the discovery of non-coding RNAs

Full-length cDNAs clones, once sequenced over the full length of the clone insert, are amenable to individual annotation in order to extrapolate their function (Kawai et al., 2001; Okazaki et al., 2002; Maeda et al., 2006; Carninci et al., 2005; Imanishi et al., 2004). Although initial attempts to annotate cDNAs were based on the assumption that all mRNAs would encode protein (Kawai et al., 2001), the expansion of the mouse cDNA collection to 61 000 cDNAs (Okazaki et al., 2002), and subsequently to 103 000 cDNAs (Maeda et al., 2006), has progressively revealed the existence of a class of generally lowly expressed transcripts apparently lacking coding potential. In fact, the discovery that in mouse there are at least 23 000 non-coding TUs came from the initial struggle to annotate these transcripts that were derived from full-length, cap-selected cDNAs, without any apparent CDS (coding sequence).

Known non-capped RNAs appear to be strongly selected against in the cap-trapped libraries. Enrichment for capped RNAs during the cap-trapping selection was calculated to be at

least 330-fold (Carninci et al., 2006). Indeed, although structural RNAs comprise more than 90% of the mammalian RNAs, examination of the raw data obtained from RIKEN 3′ ESTs (1 512 533 sequences) reveals that there are only 758 ribosomal cDNAs and 6516 mitochondrially derived cDNAs (of which 3842 were derived from only 12 problematic libraries out of 249). This proportion of cDNAs deriving from non-capped RNA is much lower than the frequency of these RNAs in cells, suggesting that these novel cDNAs, lacking coding potential, were unlikely to be genomic cDNA contamination. We further analyzed these cDNAs by computation, and identified a set of 4280 cDNAs that mapped far from existing loci, with multiple proof of their existence as *bona fide* non-coding RNAs (ncRNAs) (Numata et al., 2003). Experimental validation of novel ncRNAs that map in the mouse *Gnas* locus demonstrated the existence of eight new imprinted transcripts (Holmes et al., 2003). Further large-scale validation was performed, showing that ncRNAs are dynamically regulated in macrophages upon induction with lipopolysaccharides, further confirming that they are real RNA transcripts (Ravasi et al., 2006).

Further insights on the function of the ncRNAs derive from the observation that a large fraction of RNAs are transcribed from both orientations of the genome, thus forming sense–antisense (S/AS) transcript pairs, in which ncRNAs are often involved. These were first identified in the mouse (Okazaki et al., 2002; Kiyosawa et al., 2003) and later in human (Yelin et al., 2003). Further analysis proved that antisense ncRNAs are dynamically regulated and tend to be nuclear (Kiyosawa et al., 2005). CAGE tag data suggested that the extent of the S/AS transcription is much larger than previously estimated, by identification of bi-directional transcription for 72% of the TUs, and in particular for 86% of the TUs that map in genomic imprinted regions (loci containing genes that are expressed either paternally or maternally), suggesting that these transcripts may be involved in regulating entire complex loci (Katayama et al., 2005). The S/AS rate was further supported with 50% estimation by mouse Serial Analysis of Gene Expression (SAGE) data (Siddiqui et al., 2005). Further evidence of the regulation logic derives from the identification of over 2000 ‘chains’, or groups of transcriptional units that are overlapping or share a bidirectional promoter. These chains are to some extent conserved between mouse and human and are hypothesized to group genes under the same epigenetic regulation (Engstrom et al., 2006).

The enormous transcripts (ENEOR) consist of a group of at least 66 very large (~92 kb average) non-polyadenylated noncoding RNA, which have not been clonable with standard techniques due to size limitation of cloning vectors. These were identified by observing the presence of 3′-truncated cDNA clones primed in A-rich stretches, and reconstructing their structure by multiple RT–PCR. These ENEOR span very large regions, including various TUs, identify imprinted and micro-RNA (miRNA) genes, and may have a regulatory effect on the chromatin, as in the case of the AIR gene (Furuno et al., 2006).

The observation of ENEOR is in line with the initial analysis of 5′–3′ ditags. In fact, a large part of the cDNA population of primary lambda libraries, constituted by cDNAs longer than 6–7 kb (Carninci et al., 2002), usually does not survive large-scale propagation/sequencing operations. To overcome this, we prepared libraries containing only tags from the 5′ and 3′ ends of transcripts (Carninci et al., 2005) that were derived from large insert size cDNAs cloned in lambda FLC vectors (Carninci et al., 2001), which allows cloning of cDNAs without size bias as long as the cDNAs do not exceed 15 kbp. Large-scale sequencing of these ditags libraries suggests not only that the number of total independent transcript is larger than that identified with full-length cDNAs, but also that there are very large transcribed genomic regions called gene forests (see Glossary). Large RNAs identified by ditags span regions as large as 2 Mbp and group the TUs identified by cDNA into very large transcribed forests (Carninci et al., 2005). These 5′–3′ ditags represent borders of a part of the missing transcriptome.

The identification of non-coding RNA was initially met with scepticism, mainly because they are relatively poorly conserved between species (Wang et al., 2004; Pang et al., 2006). Despite this, their putative promoters are well conserved (Carninci et al., 2005), suggesting that their expression rather than their sequence may be biologically more important. As they may be involved in S/AS, or produce shorter RNAs (such as miRNAs), their full-length sequence conservation might indeed not be biologically relevant. For a more dedicated discussion on the function of these non-protein-coding RNAs, see (Mattick, 2003; Mehler and Mattick, 2006; Mattick and Makunin, 2006; Mattick, 2007; Carninci, 2006).

Missing transcriptome

The human transcriptome has also been extensively analyzed in the Mammalian Gene Collection (MGC) by isolating cDNAs from full-length libraries (Strausberg et al., 2002). However, these efforts greatly differ from the RIKEN approach, which is based on serial subtraction using ‘drivers’ deriving from the pools of cDNA already isolated. Instead, the MGC project has been based on outsourcing the preparation of cDNA libraries to various collaborators in at least 11 research groups (Gerhard et al., 2004), which is not compatible with serial subtraction strategies. Another key difference is the purpose of the MGC, which aims to produce at least one full-length cDNA sequence for every protein coding gene. Therefore, after sequencing the 5′ end, the clones that do not show any potential coding region are not further used for full-insert cDNA sequencing. This clearly causes under-representation of non-coding RNAs in public databases. By contrast, the selection regime for isolating full-length cDNAs used by RIKEN has been hypothesis-free: all seemingly new clones have been fully sequenced.

Human and rat transcriptomes have also been extensively sampled using subtracted/normalized ESTs from non-full-length libraries. The main difference from the RIKEN project, is that the other widespread normalization/subtraction technology (Bonaldo et al., 1996) uses double-strand cDNAs drivers. This is likely to remove antisense- as well as sense-cDNAs, thereby rendering comparisons of S/AS across different transcriptome datasets irrelevant.

The widespread existence of non-coding human RNA transcription was recently vindicated by work with whole-genome tiling arrays: upon experimental validation, some 60% of S/AS transcription rate was confirmed in the human genome (Cheng et al., 2005).

Different methodologies give rise to very great differences among datasets. In contrast with genome sequencing, where shotgun strategies are well established, it is clear that we have not yet established a universal strategy for analyzing the transcriptome, which differs from the genome in its inherent complexity. Genome sequencing alone is insufficient to compare biological phenomena because (1) comparative analysis cannot interpret a large fraction of conserved but not expressed genomic regions, (2) expressed RNAs and regulatory elements, including promoters, show different levels of conservation, and (3) low or absent conservation may be important for species-specific structural and regulatory functions. For example, the broad, CpG type of human promoters are evolutionarily more plastic, and mutate faster, than the average genomic regions in the recent human lineage, compared to the chimpanzee, in contrast to sharp, TATA-box promoters, which tend generally to be more conserved (Taylor et al., 2006; Carninci et al., 2006). Because there is such a variable degree of conservation of RNAs and regulatory elements, strategies based on genome conservation to identify genes and expressed transcripts are unacceptably hypothesis-bound.

Conversely, transcriptomics datasets are still very far from being comprehensive and comparable, due to lack of sampling, shallow sequencing, subtraction and normalization and diversification of libraries. Transcriptome analysis takes advantage of the specific interest of scientists in particular sets of expressed genes in particular tissues, but data is not systematically collected, and consequently comparison of transcriptome datasets between different organisms is inconclusive.

Even more among short RNA

With the discovery of the RNA-interference (RNAi) phenomena in *C. elegans* (Fire et al., 1998) and the discovery that these short siRNAs (~23 nt) control transcript levels in mammalian cells (Elbashir et al., 2001), the research community embarked on the medium-scale cloning and sequencing of these short (20–25 nt) RNAs. These studies surprisingly revealed the existence for the first time of microRNAs (miRNAs), which are very highly expressed (Lau et al., 2001) and regulate mRNA expression level in a large variety of biological contexts, including development, differentiation and cancer. Although sequencing costs have so far constrained the exploration of this new transcriptional world, very high-throughput sequencing methods are starting to show the full-extent of this phenomenon. For instance, in *Arabidopsis* there are more than 75 000 short RNAs (19–25 nt) (Lu et al., 2005), and a similar approach in the mouse has conservatively identified more than 20 000 sequences, among which 3374 were considered highly reliable (Mineno et al., 2006). More recently, analysis of different sized short RNAs (29–30 nt) has revealed the existence of a novel, yet-uncharacterized class of short RNAs restricted to the testis, constituted by >1000 different short RNA tags that cluster on ~100 gene-poor regions of the genome (reviewed in Kim, 2006). These short RNAs form a complex with miwi, mili or piwi RNA-binding proteins and are essential for spermatogenesis, although the exact functional mechanisms are not yet understood. Likewise, analysis of short RNAs from other tissues might soon reveal additional classes of short RNAs. These will include tissue-specific CAGE tags that identify RNA transcribed from repeated elements (G. Faulkner, K. Waki, C. O. Daub, T. Lassmann, S. Grimmond, D. Hume, Y. Hayashizaki and P. Carninci, manuscript in preparation), which could function as global genome regulators.

How many RNAs are there in a mammal?

Despite the availability of rapidly growing datasets, the true size of the transcriptome is still difficult to estimate due to the different modalities of RNA expression, their widely varying levels of expression, their compartmentalization, and the cell specificity and plasticity of RNA expression. Considering all aspects that are still underestimated (Table 1), one can envisage the existence of more than 10^6 distinguishable RNAs. However, considering all the different mammalian cell types still not explored and the considerable number of cell-specific transcripts (Kampa et al., 2004), it would not be surprising if there were at least 10^7 different mammalian transcripts. Whatever this number becomes, cells seem to produce many more RNAs than were previously recognized. How many of these RNAs are essential, redundant or dispensable, and when? This is not testable with single nucleotide mutagenesis and knock-out experiments in the laboratory, nor is it feasible to measure all of the possible phenotypes, some of which could be extremely mild, redundant or context-specific, and whose display may require different thresholds of RNA inactivation for multiple genes. Alternatively, some RNAs, such as expressed pseudogenes (Frith et al., 2006) might become functional under a new set of conditions, and thus should be considered as potential genes [or ‘potogenes’ (see Hayashizaki and Carninci, 2006)] rather than as non-genes. Finally, some of the non-coding RNAs, or groups of non-coding RNAs, might behave as genes and confer selectable traits only when a given organism is subjected to selection pressure.

The task to identify all of these different RNAs remains a substantial challenge that requires us to develop novel methodologies beyond the whole-genome tiling arrays (which cannot distinguish different overlapping transcripts and their splicing variants), the tagging technologies and individual cDNA clone analysis. Although sequencing short RNAs would fit the novel generation of sequencers developed for the $1000 genome project perfectly (Bennett et al., 2005; Margulies et al., 2005), this would not lend itself to the discovery of large (m)RNAs, because the physical combination of all splicing variants requires sequence determination of individual full-length cDNAs. Additionally, novel technologies would need to collect full-length cDNA from many more different and rare

cell types from mammalian organs, and eventually from the unexplored RNomics regions (polyA-minus and nuclear RNAs). Although the $1000 genome project might become feasible in few years, a $1000 high-resolution transcriptome is well beyond our cloning technologies due to the elusive nature of different RNA classes.

Despite these difficulties, and because comprehensive transcriptome analysis adds so much value to genome sequencing, I argue for the strategic need to standardize transcript collection methods based on comprehensive cell and condition sampling with multiple types of transcriptome libraries, combined with novel high-throughput sequencing systems. Expanding this in the comparative direction by addressing the transcriptomes of as yet unexplored organisms will surely yield biological surprises and even more novelty.

I thank all of the members of the RIKEN GSC-GREG and GSL and the Fantom-3 consortium members for data production, analysis, advice, discussions and support, and Andrew Cossins for critical reading of the manuscript. 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 (Japan).

References

Banerjee, A. K. (1980). 5′-terminal cap structure in eukaryotic messenger ribonucleic acids. *Microbiol. Rev.* **44**, 175-205.

Bennett, S. T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005). Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373-382.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S. et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246.

Bonaldo, M. F., Lennon, G. and Soares, M. B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6**, 791-806.

Carninci, P. (2006). Tagging mammalian transcription complexity. *Trends Genet.* **22**, 501-510.

Carninci, P. and Hayashizaki, Y. (1999). High-efficiency full-length cDNA cloning. *Meth. Enzymol.* **303**, 19-44.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M. et al. (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327-336.

Carninci, P., Westover, A., Nishiyama, Y., Ohsumi, T., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Schneider, C. et al. (1997). High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res.* **4**, 61-66.

Carninci, P., Nishiyama, Y., Westover, A., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. (1998). Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl. Acad. Sci. USA* **95**, 520-524.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617-1630.

Carninci, P., Shibata, Y., Hayatsu, N., Itoh, M., Shiraki, T., Hirozane, T., Watahiki, A., Shibata, K., Konno, H., Muramatsu, M. et al. (2001). Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* **77**, 79-90.

Carninci, P., Shiraki, T., Mizuno, Y., Muramatsu, M. and Hayashizaki, Y. (2002). Extra-long first-strand cDNA synthesis. *Biotechniques* **32**, 984-985.

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D. et al. (2003). Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* **13**, 1273-1289.

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

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C. et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626-635.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G. et al. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149-1154.

Das, M., Harvey, I., Chu, L. L., Sinha, M. and Pelletier, J. (2001). Full-length cDNAs: more than just reaching the ends. *Physiol. Genomics* **6**, 57-80.

De Virgilio, C., Hottiger, T., Dominguez, J., Boller, T. and Wiemken, A. (1994). The role of trehalose synthesis for the acquisition of thermotolerance in yeast. I. Genetic evidence that trehalose is a thermoprotectant. *Eur. J. Biochem.* **219**, 179-186.

Edery, I., Chu, L. L., Sonenberg, N. and Pelletier, J. (1995). An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell. Biol.* **15**, 3363-3371.

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494-498.

Engstrom, P. G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S. L., Yang, L. et al. (2006). Complex Loci in human and mouse genomes. *PLoS Genet.* **2**, e47.

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

Frith, M. C., Wilming, L. G., Forrest, A., Kawaji, H., Tan, S. L., Wahlestedt, C., Bajic, V. B., Kai, C., Kawai, J., Carninci, P. et al. (2006). Pseudo-messenger RNA: phantoms of the transcriptome. *PLoS Genet.* **2**, e23.

Furuno, M., Pang, K. C., Ninomiya, N., Fukuda, S., Frith, M. C., Bult, C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. et al. (2006). Clusters of internally-primed transcripts reveal novel long noncoding RNAs. *PLoS Genet.* **2**, e37.

Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M., Grouse, L. H., Schuler, G., Klein, S. L., Old, S., Rasooly, R., Good, P. et al. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* **14**, 2121-2127.

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

Harbers, M. and Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods* **2**, 495-502.

Hayashizaki, Y. and Carninci, P. (2006). Genome Network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genet.* **2**, e63.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W. et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**, 807-828.

Hirozane-Kishikawa, T., Shiraki, T., Waki, K., Nakamura, M., Arakawa, T., Kawai, J., Fagiolini, M., Hensch, T. K., Hayashizaki, Y. and Carninci, P. (2003). Subtraction of cap-trapped full-length cDNA libraries to select rare transcripts. *Biotechniques* **35**, 510-516, 518.

Holmes, R., Williamson, C., Peters, J., Denny, P. and Wells, C. (2003). A comprehensive transcript map of the mouse Gnas imprinted complex. *Genome Res.* **13**, 1410-1415.

Hottiger, T., De Virgilio, C., Hall, M. N., Boller, T. and Wiemken, A. (1994). The role of trehalose synthesis for the acquisition of thermotolerance in yeast. II. Physiological concentrations of trehalose increase the thermal stability of proteins in vitro. *Eur. J. Biochem.* **219**, 187-193.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M.

et al. (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2**, e162.

Jackson, D. A., Pombo, A. and Iborra, F. (2000). The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *FASEB J.* **14**, 242-254.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G. 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.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J. et al. (2005). Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566.

Kato, S., Sekine, S., Oh, S. W., Kim, N. S., Umezawa, Y., Abe, N., Yokoyama-Kobayashi, M. and Aoki, T. (1994). Construction of a human full-length cDNA bank. *Gene* **150**, 243-250.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H. et al. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685-690.

Kim, V. N. (2006). Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. *Genes Dev.* **20**, 1993-1997.

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. and Hayashizaki, Y. (2003). Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13**, 1324-1334.

Kiyosawa, H., Mise, N., Iwase, S., Hayashizaki, Y. and Abe, K. (2005). Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res.* **15**, 463-474.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M. et al. (2006). CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211-222.

Lau, N. C., Lim, L. P., Weinstein, E. G. and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858-862.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**, 239-240.

Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C. and Green, P. J. (2005). Elucidation of the small RNA component of the transcriptome. *Science* **309**, 1567-1569.

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engstrom, P. G., Lenhard, B., Aturaliya, R. N., Batalov, S., Beisel, K. W. et al. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet.* **2**, e62.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R., Beck, C., Blistain, A., Bonaldo, M., Bowers, Y., Bowles, L. et al. (1999). An encyclopedia of mouse genes. *Nat. Genet.* **21**, 191-194.

Maruyama, K. and Sugano, S. (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**, 171-174.

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* **25**, 930-939.

Mattick, J. S. (2007). A new paradigm for developmental biology. *J. Exp. Biol.* **210**, 1526-1547.

Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.* **15** Suppl. 1, R17-R29.

Mehler, M. F. and Mattick, J. S. (2006). Non-coding RNAs in the nervous system. *J. Physiol.* **575**, 333-341.

Mineno, J., Okamoto, S., Ando, T., Sato, M., Chono, H., Izu, H., Takayama, M., Asada, K., Mirochnitchenko, O., Inouye, M. et al. (2006). The expression profile of microRNAs in mouse embryos. *Nucleic Acids Res.* **34**, 1765-1771.

Miura, K. (1981). The cap structure in eukaryotic messenger RNA as a mark of a strand carrying protein information. *Adv. Biophys.* **14**, 205-238.

Mockler, T. C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S. E. and Ecker, J. R. (2005). Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**, 1-15.

Ng, P., Wei, C. L., Sung, W. K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H. et al. (2005). Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* **2**, 105-111.

Nilsson, R., Bajic, V. B., Suzuki, H., di Bernardo, D., Bjorkegren, J., Katayama, S., Reid, J. F., Sweet, M. J., Gariboldi, M., Carninci, P. et al. (2006). Transcriptional network dynamics in macrophage activation. *Genomics* **88**, 133-142.

Numata, K., Kanai, A., Saito, R., Kondo, S., Adachi, J., Wilming, L. G., Hume, D. A., Hayashizaki, Y. and Tomita, M. (2003). Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res.* **13**, 1301-1306.

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

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K. et al. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36**, 40-45.

Pang, K. C., Frith, M. C. and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1-5.

Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M. C., Gongora, M. M. 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.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D. (in press). Mammalian RNA polymerase II core promoters – insights from genome-wide studies. *Nat. Rev. Genet.*

Sasaki, N., Nagaoka, S., Itoh, M., Izawa, M., Konno, H., Carninci, P., Yoshiki, A., Kusakabe, M., Moriuchi, T., Muramatsu, M. et al. (1998). Characterization of gene expression in mouse blastocyst using single-pass sequencing of 3995 clones. *Genomics* **49**, 167-179.

Scheetz, T. E., Laffin, J. J., Berger, B., Holte, S., Baumes, S. A., Brown, R., 2nd, Chang, S., Coco, J., Conklin, J., Crouch, K. et al. (2004). High-throughput gene discovery in the rat. *Genome Res.* **14**, 733-741.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, 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. USA* **100**, 15776-15781.

Siddiqui, A. S., Khattra, J., Delaney, A. D., Zhao, Y., Astell, C., Asano, J., Babakaiff, R., Barber, S., Beland, J., Bohacec, 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. USA* **102**, 18485-18490.

Strausberg, R. L., Feingold, E. A., Grouse, L. H., Derge, J. G., Klausner, R. D., Collins, F. S., Wagner, L., Shenmen, C. M., Schuler, G. D., Altschul, S. F. et al. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA* **99**, 16899-16903.

Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Semple, C. A. (2006). Heterotachy in mammalian promoter evolution. *PLoS Genet.* **2**, e30.

Theissen, H., Etzerodt, M., Reuter, R., Schneider, C., Lottspeich, F., Argos, P., Luhrmann, R. and Philipson, L. (1986). Cloning of the human cDNA for the U1 RNA-associated 70K protein. *EMBO J.* **5**, 3209-3217.

Thill, G., Castelli, V., Pallud, S., Salanoubat, M., Wincker, P., de la Grange, P., Auboeuf, D., Schachter, V. and Weissenbach, J. (2006). ASEtrap: a biological method for speeding up the exploration of spliceomes. *Genome Res.* **16**, 776-786.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J. and Wong, G. K. (2004). Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* **431**, 1 p following 757; discussion following 757.

Watahiki, A., Waki, K., Hayatsu, N., Shiraki, T., Kondo, S., Nakamura, M., Sasaki, D., Arakawa, T., Kawai, J., Harbers, M. et al. (2004). Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *Nat. Methods* **1**, 233-239.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R. et al. (2003). Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21**, 379-386.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. and Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892-897.

The Journal of Experimental Biology 210, 1492-1496
Published by The Company of Biologists 2007
doi:10.1242/jeb.02783

Glossary of terms

This section is designed to help readers adapt to the complex terminology associated with contemporary molecular genetics, genomics and systems biology. Fuller descriptions of these terms are available at http://www.wikipedia.org/

Ab initio prediction	methods used to predict the potential genes encoded in the genome, which are trained on datasets made of known genes, and used computationally to predict coding regions out of genome without the aid of cDNA sequence. Although their performance is improving, these algorithms perform very poorly on non-protein coding genes.
Annotation	as applied to proteins, DNA sequences or genes. The storage of data describing these entities (protein/gene identities, DNA motifs, gene ontology categorisation, etc.) within a biological database. Active projects include FlyBase and WormBase. See Gene ontology.
Assembly	the process of aligning sequenced fragments of DNA into their correct positions within the chromosome or transcript.
cDNA	complementary DNA. This is DNA synthesised from a mature mRNA template by the enzyme reverse transcriptase. cDNA is frequently used as an early part of gene cloning procedures, since it is more robust and less subject to degradation than the mRNA itself.
ChIP	chromatin immunoprecipitation assay used to determine which segments of genomic DNA are bound to chromatin proteins, mainly including transcription factors.
Chip	see Microarray.
ChIP-on-chip	use of a DNA microarray to analyse the DNA generated from chromatin immunoprecipitation experiments (see ChIP).
cis-acting	a molecule is described as *cis*-acting when it affects other genes that are physically adjacent, on the same chromosome, or are genetically linked or in close proximity (for mRNA expression, typically a promoter).
Collision-induced dissociation	a mechanism by which molecules (e.g. proteins) are fragmented to form molecular ions in the gas phase. These fragments are then analysed within a mass spectrometer to provide mass determination.
Connectivity	a term from graph theory, which indicates the number of connections between nodes or vertices in a network. Greater connectedness between nodes is generally used as a measure of robustness of a network.
CpG islands	regions that show high density of ‘C followed by G’ dinucleotides and are generally associated with promoter elements; in particular, stretches of DNA of at least 200 bp with a C–G content of 50% and an observed CpG/expected CpG in excess of 0.6. The cytosine residues can be methylated, generally to repress transcription, while demethylated CpGs are a hallmark of transcription. CpG dinucleotides are under-represented outside regulatory regions, such as promoters, because methylated C mutates into T by deamination.
Edge	as in networks. Connects two nodes (or vertices) within a system. These concepts arise from graph theory.
Enhancer	a short segment of genomic DNA that may be located remotely and that, on binding particular proteins (*trans*-acting factors), increases the rate of transcription of a specific gene or gene cluster.
Epistasis	a phenomenon when the properties of one gene are modified by one or more genes at other loci. Otherwise known as a genetic interaction, but epistasis refers to the statistical properties of the phenomenon.

eQTL	the combination of conventional QTL analysis with gene expression profiling, typically using microarrays. eQTLs describe regulatory elements controlling the expression of genes involved in specific traits.
EST	expressed sequence tag. A short DNA sequence determined for a cloned cDNA representing portions of an expressed gene. The sequence is generally several hundred base pairs from one or both ends of the cloned insert.
Exaptation	a biological adaptation where the current function is not that which was originally evolved. Thus, the defining (derived) function might replace or persist with the earlier, evolved adaptation.
Exon	any region of DNA that is transcribed to the final (spliced) mRNA molecule. Exons interleave with segments of non-coding DNA (introns) that are removed (spliced out) during processing after transcription.
Gene forests	genomic regions for which RNA transcripts, produced from either DNA strand, have been identified without gaps (non-transcribed genomic regions). Conversely, regions in which no transcripts have ever been detected are called ‘gene deserts’.
Gene interaction network	a network of functional interactions between genes. Functional interactions can be inferred from many different data types, including protein–protein interactions, genetic interactions, co-expression relationships, the co-inheritance of genes across genomes and the arrangement of genes in bacterial genomes. The interactions can be represented using network diagrams, with lines connecting the interacting elements, and can be modelled using differential equations.
Gene ontology (GO)	an ontology is a controlled vocabulary of terms that have logical relationships with each other and that are amenable to computerised manipulation. The Gene Ontology project has devised terms in three domains: biological process, molecular function and cell compartment. Each gene or DNA sequence can be associated with these annotation terms from each domain, and this enables analysis of microarray data on groups of genes based on descriptive terms so provided. See http://www.geneontology.org
Gene set enrichment analysis	a computational method that determines whether a defined set of genes, usually based on their common involvement in a biological process, shows statistically significant differences in transcript expression between two biological states.
Gene silencing	the switching-off of a gene by an epigenetic mechanism at the transcriptional or post-transcriptional levels. Includes the mechanism of RNAi.
Genetic interaction (network)	a genetic interaction between two genes occurs when the phenotypic consequences of a mutation in one gene are modified by the mutational status at a second locus. Genetic interactions can be aggravating (enhancing) or alleviating (suppressing). To date, most high-throughput studies have focussed on systematically identifying synthetic lethal or sick (aggravating) interactions, which can then be visualised as a network of functional interactions (edges) between genes (nodes).
Genome	a portmanteau of gene and chromosome, the entire hereditary information for an organism that is embedded in the DNA (or, for some viruses, in RNA). Includes protein-coding and non-coding sequences.
Heritability	phenotypic variation within a population is attributable to the genetic variation between individuals and to environmental factors. Heritability is the proportion due to genetic variation usually expressed as a percentage.
Heterologous hybridization	the use of a cDNA or oligonucleotide microarray of probes designed for one species with target cRNA/cDNAs from a different species.
Homeotic	the transformation of one body part to another due to mutation of specific developmentally related genes, notably the *Hox* genes in animals and *MADS-box* genes in plants.
Hub	as in networks. A node with high connectivity, and thus which interacts with many other nodes in the network. A hub protein interacts with many other proteins in a cell.

Hybridisation	the process of joining (annealing) two complementary single-stranded DNAs into a single double-stranded molecule. In microarray analysis, the target RNA/DNA from the subject under investigation is denatured and hybridised to probes that are immobilised on a solid phase (i.e. glass microscope slide).
Hypomorph	in genetics, a loss-of-function mutation in a gene, but which shows only a partial reduction in the activity it influences rather than a complete loss (cf. hypermorph, antimorph, neomorph, etc).
Imprinting	a phenomenon where two inherited copies of a gene are regulated in opposite ways, one being expressed and the other being repressed.
Indel	insertion and deletion of DNA, referring to two types of genetic mutation. To be distinguished from a 'point mutation', which refers to the substitution of a single base.
Interactome	a more or less comprehensive set of interactions between elements within cells. Usually applied to genes or proteins as defined by transcriptomic, proteomic or protein–protein interaction data.
Intron	see Exon.
KEGG	The Kyoto Encyclopedia of Genes and Genomes is a database of metabolic and other pathways collected from a variety of organisms. See http://www.genome.jp/kegg
Metabolomics	the systematic qualitative and quantitative analysis of small chemical metabolite profiles. The metabolome represents the collection of metabolites within a biological sample.
Metagenomics	the application of genomic techniques to characterise complex communities of microbial organisms obtained directly from environmental samples. Typically, genomic tags are sequence characterised as markers of each species to inform on the range and abundance of species in the community.
Microarray	an arrayed set of probes for detecting molecularly specific analytes or targets. Typically, the probes are composed of DNA segments that are immobilised onto the solid surface, each of which can hybridise with a specific DNA present in the target preparation. DNA microarrays are used for profiling of gene transcripts.
Model species	a species used to study particular biological phenomena, the outcome offering insights into the workings of other species. Usually, the selection is based on experimental tractability, particularly ease of genetic manipulation. For the geneticist, it is an organism with inbred lines where sibs will be >98% identical (i.e. *Drosophila*, *Caenorhabditis elegans* and mice). For genomic science, it refers to a species for which the genomic DNA has been sequenced.
miRNA	a category of novel, very short, non-coding RNAs, generated by the cleavage of larger precursors (pri-miRNA). These short RNAs are included in the RNA-induced silencing complex (RISC) and pair to the 3′ ends of target RNA, blocking its translation into proteins (in animals) or promoting RNA cleavage and degradation (in plants).
mRNA	a protein-coding mRNA containing a protein-coding region (CDS), preceded by a 5′ and followed by a 3′ untranslated region (5′ UTR and 3′ UTR). The UTRs contain regulatory elements. A full-length cDNA contains the complete sequence of the original mRNA, including both UTRs. However, it is often difficult to assign the starting–termination positions for protein synthesis unambiguously. A cDNA containing the entire CDS is often considered acceptable for bioinformatic and experimental studies requiring full-length cDNAs.
ncRNA	non-coding RNA is any RNA molecule with no obvious protein-coding potential for at least 80 or 100 amino acids, as determined by scanning full-length cDNA sequences. It includes ribosomal (rRNA) and transfer RNAs (tRNA) and is now known to include various sub-classes of RNA, including snoRNA, siRNA and piRNA. Just like the coding mRNAs, a large proportion of ncRNAs are transcribed by RNA polymerase II and are large transcripts. A description of the many forms of ncRNA can be found at http://en.wikipedia.org/wiki/Non-coding_RNA.

Node	as in networks. Objects linked by edges to create a network.
PCR	polymerase chain reaction. A molecular biology technique for replicating DNA *in vitro*. The DNA is thus amplified, sometimes from very small amounts. PCR can be adapted to perform a wide variety of genetic manipulations.
piRNA	Piwi-interacting RNA. A class of RNA molecules (29–30 nt long) that complex with Piwi proteins (a class of the Argonaute family of proteins) and are involved in transcriptional gene silencing.
PMF	peptide mass fingerprinting. An analytical technique for protein identification in which a protein is fragmented using proteases. The resulting peptides are analysed by mass spectrometry and these masses compared against a database of predicted or measured masses to generate a protein identity.
Polyadenylation	the covalent addition of multiple A bases to the 3′ tail of an mRNA molecule. This occurs during the processing of transcripts to form the mature, spliced molecule and is important for regulation of turnover, trafficking and translation.
Post-source decay	in mass spectrometry. The fragmentation of precursor molecular ions as they accelerate away from the ionisation source of the mass spectrometer. All precursor ions leaving the ion source have approximately the same kinetic energy, but fragmentation results in smaller product ions that can be distinguished from precursor ions using a ‘reflectron’ by virtue of their lower kinetic energies.
Post-translational modification	the chemical modification of a protein after synthesis through translation. Some modifications, notably phosphorylation, affect the properties of the protein, offering a means of regulating function.
Principal component analysis (PCA)	a technique for simplifying complex, multi-dimensional datasets to a reduced number of dimensions, the principal components. This procedure retains those characteristics of the data that relate to its variance.
Promoter	a regulatory DNA sequence, generally lying upstream of an expressed gene, which in concert with other often distant regulatory elements directs the transcription of a given gene.
Proteome	the entire protein complement of an organism, tissue or cell culture at a given time.
Quantitative trait	inheritance of a phenotypic property or characteristic that varies continuously between extreme states and can be attributed to interactions between multiple genes and their environment.
qPCR	quantitative real-time PCR, sometimes called real-time PCR. A more quantitative form of RT-PCR in which the quantity of amplified product is estimated after each round of amplification.
QTL	quantitative trait loci. A region of DNA that contains those genes contributing to the trait under study.
RISC	RNA-induced silencing complex. A protein complex that mediates the double-stranded RNA-induced destruction of homologous mRNA.
RNAi	RNA interference or RNA-mediated interference. The process by which double-stranded RNA triggers the destruction of homologous mRNA in eukaryotic cells by the RISC.
RT-PCR	reverse transcription–polymerase chain reaction. A technique for amplifying a defined piece of RNA that has been converted to its complementary DNA form by the enzyme reverse transcriptase. See qPCR.
siRNA	small interfering RNA, or silencing RNA. A class of short (20–25 nt), double-stranded RNA molecules. It is involved in the RNA interference pathway, which alters RNA stability and thus affects RNA concentration and thereby suppresses the normal expression of specific genes. Widely used in biomedical research to ablate specific genes.

snoRNA	small nucleolar RNA. A sub-class of RNA molecules involved in guiding chemical modification of ribosomal RNA and other RNA genes as part of the regulation of gene expression.
SNP	single nucleotide polymorphism. A single base-pair mutation at a specific locus, usually consisting of two alleles. Because SNPs are conserved over evolution, they are frequently used in QTL analysis and in association studies in place of microsatellites, and in genetic fingerprinting analyses.
SSH	suppressive subtractive hybridisation. A powerful protocol for enriching cDNA libraries for genes that differ in representation between two or more conditions. It combines normalisation and subtraction in a single procedure and allows the detection of low-abundance, differentially expressed transcripts, such as those involved in signalling and signal transduction.
Structural RNAs	a class of non-coding RNA, long known to have a structural role (for instance, the ribosomal RNAs), transcribed by RNA polymerase I or III.
Systems biology	treatment of biological entities as systems composed of defined elements interacting in defined ways to enable the observed function and behaviour of that system. The properties of the systems are embedded in a quantitative model that guides further tests of systems behaviour.
TATA-boxes	sequences in promoter regions constituted by TATAAA, or similar variants, which were considered the hallmark of Promoters. Recent data show that they are present only in the minority of promoters, where they direct transcription at a single well-defined location some 30 bp downstream of this element.
trans-acting	a factor or gene that acts on another unlinked gene, a gene on a separate chromosome or genetically unlinked usually through some diffusible protein product (for mRNA expression, typically a transcription factor).
Transcript	an RNA product produced by the action of RNA polymerase reading the sequence of bases in the genomic DNA. Originally limited to protein-coding sequences with flanking UTRs but now known to include large numbers of products that do not code for a protein product.
Transcriptome	the full set of mRNA molecules (transcripts) produced by the system under observation. Whilst the genome is fixed for a given organism, the transcriptome varies with context (i.e. tissue source, ontogeny, external conditions or experimental treatment).
Transgene	a gene or genetic material that has been transferred between species or between organisms using one of several genetic engineering techniques.
Transinduction	generation of transcripts from intergenic regions. At least some such products do not relate to a definable promoter or transcriptional start site.
Transposon	sequences of DNA able to move to new positions within the genome of a single cell. This event might cause mutation at the site of insertion. Also called ‘mobile genetic elements’ or ‘jumping genes’.
Transvection	an epigenetic phenomenon arising from the interaction between one allele and the corresponding allele on the homologous chromosome, leading to gene regulation.
TUs	transcriptional units. Used to group all of the overlapping RNA transcripts that are transcribed from the same genomic strand and share exonic sequences.
UTR	untranslated region. Regions of the mRNA that lie at either the 3′ or 5′ flanking ends of the molecule (i.e. 3′ UTR and 5′ UTR). They bracket the protein-coding region and contain signals and binding sites that are important for the regulation of both protein translation and RNA degradation.