Functional transcriptomics in the post-ENCODE era

Jonathan M. Mudge,[1] Adam Frankish, and Jennifer Harrow

Department of Informatics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom

The last decade has seen tremendous effort committed to the annotation of the human genome sequence, most notably perhaps in the form of the ENCODE project. One of the major findings of ENCODE, and other genome analysis projects, is that the human transcriptome is far larger and more complex than previously thought. This complexity manifests, for example, as alternative splicing within protein-coding genes, as well as in the discovery of thousands of long noncoding RNAs. It is also possible that significant numbers of human transcripts have not yet been described by annotation projects, while existing transcript models are frequently incomplete. The question as to what proportion of this complexity is truly functional remains open, however, and this ambiguity presents a serious challenge to genome scientists. In this article, we will discuss the current state of human transcriptome annotation, drawing on our experience gained in generating the GENCODE gene annotation set. We highlight the gaps in our knowledge of transcript functionality that remain, and consider the potential computational and experimental strategies that can be used to help close them. We propose that an understanding of the true overlap between transcriptional complexity and functionality will not be gained in the short term. However, significant steps toward obtaining this knowledge can now be taken by using an integrated strategy, combining all of the experimental resources at our disposal.

Over one hundred years after the basic rules of heredity were established, the gene is undergoing an identity crisis. Indeed the question "what is a gene?" has been much debated in recent years (Mattick 2003; Pearson 2006; Gerstein et al. 2007; Gingeras 2007; Pennisi 2007; Brosius 2009; Mercer and Mattick 2013). In a scientific context, this question concerns the way in which information is stored in the genome. Over the 20th Century, the biological definition of the gene evolved from "the site of a hereditable trait" to "the genomic region from where the mRNA that encodes a protein is transcribed," i.e., the "central dogma" of molecular biology (Fig. 1A; Crick 1970). In the 21st Century, however, our view of transcription is becoming more complicated. In particular, a locus may generate multiple transcripts due to alternative splicing (AS) (Harrow et al. 2012) and read-through transcription (Fig. 1B; Frenkel-Morgenstern et al. 2012), while the discovery of long noncoding RNAs (lncRNAs) suggests that most human transcripts may not encode proteins (Rinn and Chang 2012). In fact, the bulk of the genome appears to be "pervasively" transcribed (The ENCODE Project Consortium 2012), although the functional relevance of this process remains a source of debate (Ball 2013; Doolittle 2013; Graur et al. 2013). We use the term "transcriptional complexity" to refer to these phenomena collectively.

This complexity complicates the work of scientists tasked with describing the human genome. Furthermore, the 20th Century concept of the gene has become ingrained in wider society, for example as part of the language in which scientists discuss their work with laypeople and clinicians talk to patients. The modern effort to redefine the gene is thus a practical endeavor, attempting to "retrofit" biological complexity into an existing vocabulary such that it remains workable for scientists across a range of disciplines. To this end, Gerstein and colleagues recently proposed that "a gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products" (Gerstein et al. 2007). The key point here is that the word "gene" no longer designates a unit of functionality. Instead it is used as a collective term for a group of products, i.e., transcripts. From our perspective, there are vital questions concealed within the "what is a gene?" debate. For example: what is the true size of the transcriptome and what proportion of this transcription is genuinely functional? Indeed, what does "functional" actually mean in this context? Here, we summarize our current knowledge on these issues, highlight the pressing need to close those gaps in our knowledge that remain, and discuss the ways in which functionality can be captured in gene annotation.

The capture and annotation of transcript models

Clearly, transcript capture precedes transcript annotation. Previously, transcripts were captured as cDNAs, mRNAs, and ESTs, whereas today RNA-seq methodologies predominate given their high-throughput nature (Mortazavi et al. 2008; Pan et al. 2008; Wang et al. 2008, 2009; Robertson et al. 2010; Martin and Wang 2011; Ozsolak and Milos 2011; Gonzalez-Porta et al. 2012). However, gene annotation projects do not simply capture transcripts; they also provide a prediction into their biological function. There are several large-scale gene annotation projects in progress on the human genome, including RefSeq (Pruitt et al. 2005), GENCODE (Harrow et al. 2012), and UCSC Genes (Dreszer et al. 2012). In each "gene set" or "genebuild" produced, the vast majority of models are based upon transcriptomics data. Briefly, GENCODE (the gene set of the ENCODE project [The ENCODE Project Consortium 2012]) represents a merge between manually annotated HAVANA and computationally derived Ensembl models, with annotation taking place on the genome sequence. In contrast, while RefSeq also combines manual and automated processes, most human annotation takes place on full-length cDNAs that are subsequently linked to the chromosome. Finally, UCSC Genes combine RefSeq models mapped to the genome with additional models from other data sources, for example computational models based on GenBank ESTs.

[1] **Corresponding author**
E-mail jm12@sanger.ac.uk

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.161315.113. Freely available online through the *Genome Research* Open Access option.

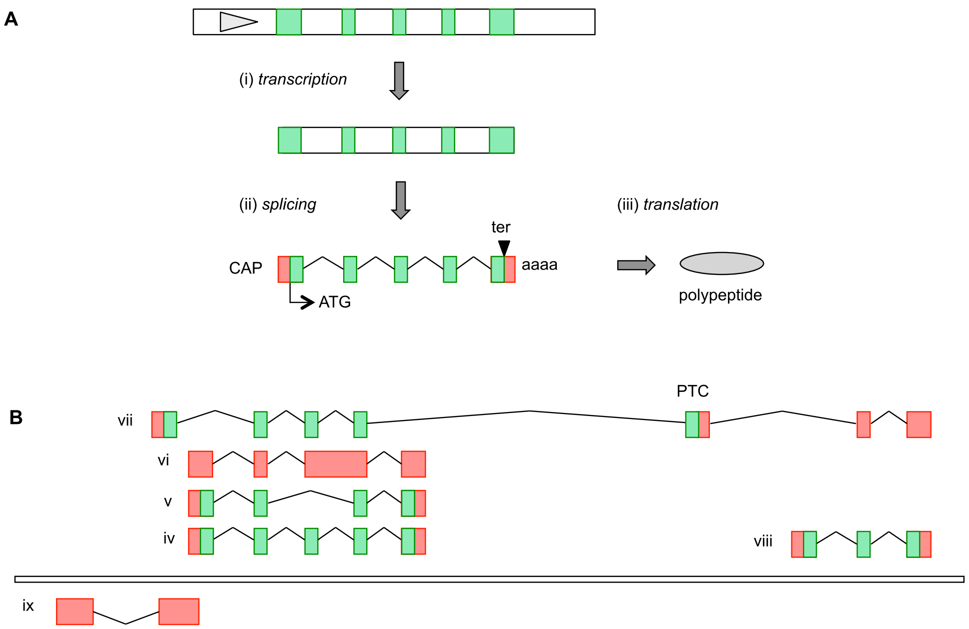

Figure 1. The evolving dogma of gene transcription. (*A*) The historical "central dogma" of molecular biology. By this model, (i) transcription generates the primary transcript (exons in green, introns in white), with the initial interaction between the RNA polymerase complex and the genome being mediated by a promoter region (gray triangle). (ii) The introns of the primary transcript are removed by the spliceosome, and a mature mRNA is generated by 5′ end capping (CAP) and polyadenylation (aaaa) (coding region [CDS] shown in green, untranslated 5′ and 3′ UTRs in red). (iii) The mRNA is translated into a polypeptide by the ribosome complex, with translation proceeding from the initiation codon (ATG) and ending at the termination codon (ter). (*B*) An updated model reflecting a modern view of transcriptional complexity. Here, the same gene (iv) undergoes alternative splicing (AS), for example an exon skipping event that does not change the frame of the CDS (v); this event thus has the potential to generate an alternative protein isoform. However, products of AS cannot be assumed to be functional; this gene has generated a retained intron transcript (vi), perhaps due to the failure of the spliceosome to remove this intron. Further complexity comes from a read-through transcription event (vii), whereby a transcript is generated that also includes exons from a neighboring protein-coding locus (viii). In this example, the read-through transcript has an alternative first exon compared with the upstream gene that contains a potential alternative ATG codon, although the presence of a subsequent premature termination codon (PTC) prior to two splice junctions indicates that this transcript is likely subjected to the nonsense mediated decay (NMD) degradation pathway. Finally, model ix is a transcript that is antisense to the upstream gene; both loci are potentially generated under the control of a bidirectional promoter.

The remit of GENCODE is to capture all nonredundant transcripts as models, and we will therefore use this gene set to discuss transcriptional complexity. GENCODE version 16 contains 194,034 transcripts found within 56,563 genes (Harrow et al. 2012), although less than half of these genes are protein-coding, as summarized in Figure 2. LncRNA is an umbrella term for transcripts that are not associated with protein-coding loci, with a minimum size of 200 bp typically used to distinguish them from small RNAs (The FANTOM Consortium et al. 2005; Kapranov et al. 2007; Guttman et al. 2009; Clark and Mattick 2011; Mattick 2011; Wang and Chang 2011; Guttman and Rinn 2012; Moran et al. 2012; Rinn and Chang 2012). The 13,220 lncRNA loci in GENCODE v16 are subcategorized according to their spatial relationship with protein-coding genes. Pseudogenes are classified according to their mode of formation: either by mRNA retroinsertion (processed pseudogenes), gene duplication (unprocessed pseudogenes), or by the inactivation of functional genes (unitary pseudogenes). While pseudogenes do not contain an intact or translated CDS, at least 9% of human pseudogenes are transcribed (Pei et al. 2012), and there is evidence that pseudogene loci can gain new functionality via "resurrection" (Brosch et al. 2011; Pei et al. 2012; Johnsson et al. 2013). Finally, GENCODE contains numerous categories of small RNA, which are beyond the scope of this article. These include rRNA and tRNA loci, which are believed to be well described (International Human Genome Sequencing Consortium 2001; Uechi et al. 2001; Flicek et al. 2013), as well as more recently discovered categories including microRNAs (Kozomara and Griffiths-Jones 2011) and piwi-interacting RNAs (Siomi et al. 2011). Small RNA genes are frequently found within the exons or introns of larger transcripts (both protein-coding and lncRNA), which can be regarded as "host" transcripts (Djebali et al. 2012a).

Toward a definition of transcript functionality

What is a functional transcript? An RNA that is translated into protein is clearly functional. However, this is not the only mode by which transcripts can influence physiology. Consider the nonsense-mediated decay (NMD) pathway, which degrades transcripts featuring premature termination codons (PTCs) (Fig. 1B, model vii; Mendell et al. 2004). While a major role of NMD is to "mop up" transcriptional errors, certain genes utilize NMD for gene regulation (Lareau et al. 2007; Huang et al. 2011). Such genes can switch transcription from a CDS transcript to an NMD-targeted transcript in order to reduce protein output. Is the NMD transcript functional? It clearly does not function in the same way that a protein-coding transcript does. Nonetheless, the act of its creation imparts functionality, i.e., gene regulation. Consider also the *Airn* lncRNA locus in mouse, which induces the imprinted silencing of the *Igf2r* gene found on the opposite strand (Sleutels et al. 2002). By shortening the endogenous lncRNA transcript, Latos and colleagues recently demonstrated that it is the act of *Airn* transcription that drives *Igf2r* silencing; the *Airn* transcript appears to be a by-product of this process (Latos et al. 2012). It may therefore be helpful in regard to the concept of a functional transcript as distinct from the

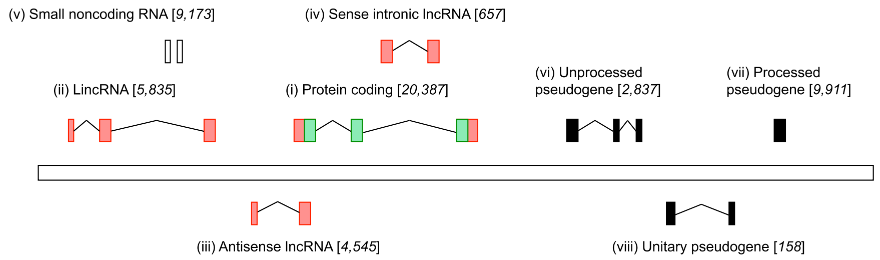
(v) Small noncoding RNA [9,173]
(iv) Sense intronic lncRNA [657]
(ii) LincRNA [5,835]
(i) Protein coding [20,387]
(vi) Unprocessed pseudogene [2,837]
(vii) Processed pseudogene [9,911]
(iii) Antisense lncRNA [4,545]
(viii) Unitary pseudogene [158]

Figure 2. A summary of locus biotypes in GENCODE. This schematic details the major classes of loci found in the GENCODE v16 human gene set, and in square brackets the total number of each set. These counts are made at the locus level as opposed to the transcript level. GENCODE contains 194,034 transcripts in total, 81,626 of which have an annotated CDS. This means there is an average of 4.0 CDS transcripts per protein-coding gene, while 14,786 protein-coding genes contain more than one distinct CDS (i). Long intergenic RNAs (lincRNAs), antisense RNAs, and sense intronic RNAs are treated as subbiotypes of lncRNA (ii–iv). In GENCODE, lincRNAs are models that do not overlap a protein-coding gene or pseudogene on either strand, antisense RNAs are models found on the opposite strand to exons or introns of protein-coding genes, and sense intronic RNAs are found entirely with the intron of a protein-coding gene. In total, GENCODE contains 22,444 lncRNA transcripts, an average of 1.7 per lncRNA locus. (v) The 9173 loci classed as small noncoding RNA loci include the classic rRNA and tRNA genes, as well as the more recently identified categories of loci such as miRNAs, snoRNAs, and piRNAs. The 13,419 pseudogenes found in GENCODE can be divided into three major classes: unprocessed, processed, and unitary (vi–viii). Unprocessed pseudogenes result from the genomic duplication of protein-coding genes; pseudogenization may come from the fact that the duplication is partial, or by subsequent mutation. Processed pseudogenes are formed by the retroinsertion of mRNAs into the genome sequence, and these loci are thus typically intronless. Unitary pseudogenes are protein-coding genes that are pseudogenized in the human lineage, as judged by a comparison with an intact coding ortholog in another species. Further to this diagram, GENCODE also contains 26 polymorphic pseudogenes: models in the reference assembly that are known to exist as intact protein-coding loci in other human genomes. All classes of pseudogenes may be subjected to transcription.

concept of functional transcription (Kornienko et al. 2013). However, in the context of annotation, we believe it is appropriate to define a functional transcript as one that makes a contribution to phenotypic complexity, regardless of the mechanism by which this occurs.

What, then, is a nonfunctional transcript? First, we define nonfunctional transcripts as those created by biological mechanisms as opposed to technical artifacts of the experimental process, e.g., genomic DNA contaminants ("artifact" transcripts can be common in RNA libraries, and it is vitally important that they are filtered out of annotation projects [Harrow et al. 2012]). Fundamentally, gene expression is a stochastic process, whereby variations in its output arise from the random nature of the underlying molecular interactions (Raser and O'Shea 2005; Munsky et al. 2012). Stochastic effects can occur during transcription and splicing, making both processes a potential source of "incorrect" reactions. If the resulting transcripts are biologically inert they could be regarded as "noise" (Melamud and Moult 2009). For example, while the spliceosome is believed to be highly accurate it does not act with complete fidelity (Hsu and Hertel 2009). GENCODE contains 25,466 models classed as retained introns (Fig. 1B, model vi), and we suspect that many result from the failure of the spliceosome to initiate or complete the splicing of a particular intron. In fact, the genome sequence motifs that govern splicing are commonly suboptimal, increasing the likelihood that "correct" splice sites will be missed by the spliceosome. As well as intron retention, this can lead to exon "skips" (Fig. 1B, model v) and the utilization of alternative de novo or "cryptic" splice sites (Pickrell et al. 2010b).

Second, it is noteworthy that 62% of lncRNA transcripts in GENCODE overlap transposable elements (TEs) (JM Mudge, unpubl.). Furthermore, there is a well-established link between TEs and AS in protein-coding genes; certain TE families such as *Alu* contain DNA motifs that resemble splicing signals, making them a ready substrate for exon creation events ("exonization") (Sorek 2007; Shen et al. 2011). Since detectable TE insertions are typically seen to have occurred during recent evolution, it appears that most TE transcripts or exons are also young (Sorek 2007). These observations have led to TEs being considered a significant source of noise. Even so, the number of TE transcripts known to be functional is increasing, indicating that TE sequences should also be regarded as a potential source of evolutionary innovation (Camacho-Vanegas et al. 2012; Zarnack et al. 2013). Consider also the possibility that TEs may insert into existing transcripts without disrupting functionality. Most obviously, the 3′ untranslated regions (UTRs) of protein-coding transcripts in GENCODE are replete with TEs (JM Mudge, unpubl.). Functional transcripts may thus contain "nonfunctional" sequence. In fact, a "functional vs. nonfunctional" model is likely to be a false dichotomy in practice. If we assume that transcript creation (via TE insertion or de novo mutation) is the first step toward the generation of new functionality in the transcriptome, then we should anticipate the existence of transcripts that are in the process of being "tested" for functionality (Modrek and Lee 2003; Brosius 2005). This suggests that there may not always be a watershed separating transcriptional noise from evolutionary novelty, and that functionality is in reality an analog as opposed to a binary classification. Certain transcripts may thus possess "minor" functionality, perhaps making a contribution to cellular physiology while remaining inessential for overall viability. Indeed, "functional" clearly does not mean "essential to survival"; many functional human genes (such as olfactory receptors) can be found pseudogenized in the genomes of healthy adults (these are referred to as "loss-of-function" [LoF] genes) (MacArthur et al. 2012).

The importance of classifying functionality in the transcriptome

There is a conceptual difference between demonstrating that a transcript is functional and describing what that function actually is. Most obviously, the majority of the 20,387 protein-coding genes in GENCODE contain at least one CDS transcript considered "known," meaning it generates a protein molecule recognized in the manually curated UniProt database (The UniProt Consortium 2012). Nonetheless the majority of the 81,626 GENCODE CDSs lack experimental support for translation (Harrow et al. 2012). Furthermore, the proportion of lncRNA models that have been confirmed to function as noncoding transcripts is minute (see Toward a Functional Annotation of lncRNAs section). This means that, while GENCODE is a larger gene set than RefSeq, it contains a higher proportion of transcripts of putative functionality. Such gaps in our knowledge can have significant effects on scientific analyses; effects that we believe remain largely unappreciated. Consider a biologist who is investigating genome sequence variants found within a protein-coding gene. If she chooses to work with a single experimentally confirmed coding transcript, the effect of variants found within this CDS can be interpreted with some confidence. However, if she considers all transcripts within the locus, this may completely recontextualize the functional interpretation of the variant site. Furthermore, she may identify additional variants that overlap with exonic sequence. In a recent study by MacArthur and colleagues, a third of sequence variants predicted to cause LoF in human genes were found to be subject to AS, including one in *ZSCAN9* (Fig. 3; MacArthur et al. 2012). If only the model containing the variant had been considered (D), it would appear as if the entire gene were subjected to LoF. In fact, the authors could see that *ZSACN9* also contains three CDSs (A, B, and C) that do not contain this variant. LoF can now be seen as an attribute of transcript D, and not necessarily of the whole gene. However, the accuracy of this interpretation depends on our confidence in the functional annotation of the locus. If the AS event that incorporates the variant site into transcript D is spurious, then the variant may not be of biological interest; *ZSCAN9* would not be a LoF gene. Conversely, if transcript D is functional though A, B, and C are not, a prediction that this gene escapes LoF due to AS would be unfounded. In fact, we are not aware of any laboratory data that unambiguously confirms the existence of any of the four predicted *ZSACN9* protein molecules in vivo.

So where do we go from here? The importance of generating a complete functional annotation of the human transcriptome cannot be understated. Unfortunately, neither can the difficulty. Ultimately, when describing functionality, there is no substitute for the detailed experimental dissection of a single gene as performed in the laboratory. Clearly, such work will continue for the foreseeable future, both for protein-coding genes and lncRNAs. The downside to this approach is that it is time consuming, and the techniques commonly used cannot be readily scaled up to examine large numbers of transcripts in a single study. It is therefore inevitable that scientists have explored the usefulness of whole-transcriptome strategies in the annotation of functionality, essentially by combining a wide variety of modern high-throughput techniques with the power of computational biology. We believe that such approaches can make significant progress in this regard, in spite of their potential limitations. In fact, improvements to gene annotation will be of enormous benefit to single gene studies. For example, if scientists have knowledge of the different transcripts that a gene produces, alongside insights into their potential functionality, then this will aid the design of assays that are specific to a certain transcript within that gene.

New technologies can aid transcript capture and completion

How many human transcripts remain to be discovered? To begin with, we must recognize that transcriptomes can differ significantly between the cells of distinct tissues and developmental stages, in terms of both the transcripts produced and their levels of expression (Heinzen et al. 2008; Pan et al. 2008; Wang et al. 2008; Brawand et al. 2011; Kang et al. 2011). Furthermore, splicing abnormalities are commonly observed in cancer cells and immortalized cell lines (a source of much ENCODE transcriptomic data) (Brinkman 2004; Wang and Cooper 2007; Chen et al. 2011; Djebali et al. 2012a; The ENCODE Project Consortium 2012). Finally, it is becoming apparent that AS patterns can show notable polymorphism (Montgomery et al. 2010; Pickrell et al. 2010a; Gonzalez-Porta et al. 2012). While there may be no single human transcriptome, it will often make practical sense to work with a “consensus” transcriptome that combines all known transcripts into one gene set. Even then, the question of missing transcripts is difficult to answer. First, any of the million exons in GENCODE could theoretically be subject to a variety of splicing errors. This suggests the set of transcripts that can be detected experimentally is infinitely large, and that we should perhaps count only those transcripts that show consistently reproducible expression (see The Annotation of Gene Expression Data section). In this way, a recent RT-PCR analysis suggests that a fifth of GENCODE genes contain exons that have yet to be annotated (Howald et al. 2012).

Second, we must consider pervasive transcription. ENCODE found that 62.1% of the genome (combined across 15 cell lines) is covered by processed transcripts extrapolated from sequencing reads (Djebali et al. 2012a), with 34% of the bases incorporated being intergenic. Other studies have since reported similar findings (Hangauer et al. 2013). Furthermore, it has been shown that transcription proceeds in the 5′ direction beyond 65% of ENCODE gene boundaries, often integrating into the exonic structure of upstream genes; the role of such “chimeric read-through RNAs” remains largely unclear (Fig. 1B, model vii; Gingeras 2009; Djebali et al. 2012b; Frenkel-Morgenstern et al. 2012). Unfortunately, contemporary RNA-seq data sets remain a source of technical frustration; the reads generated are short in length, and it is no trivial task to assemble these fragments into full-length transcript models (Wang et al. 2009; Martin and Wang 2011; Ozsolak and Milos 2011). Furthermore, technical differences amongst the variety of sequencing methodologies available can lead to variations in the nature and quality of the sequences obtained. For example, protocols that incorporate a PCR-based amplification step may show a bias toward the capture of highly expressed transcripts (Sam et al. 2011); in fact, the amplification step can be a source of experimental artifact (Mamanova et al. 2010). Second, RNA sample preparations commonly incorporate selection for polyadenylated RNAs, chiefly to avoid capturing rRNA. However, the cell contains a large amount of non-polyadenylated RNA that is not rRNA, and these poorly understood transcripts will therefore be lost if this filtering step is used (Raz et al. 2011; Djebali et al. 2012a; Livyatan et al.

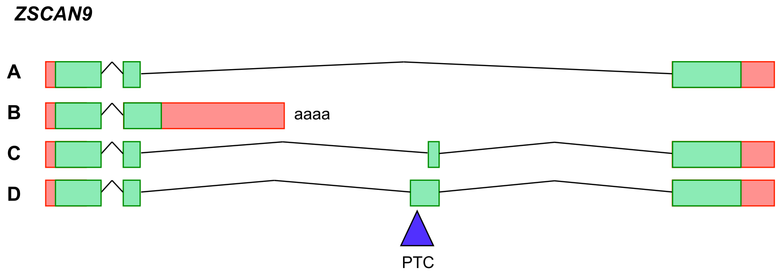

Figure 3. A LoF variant within the zinc finger and SCAN domain containing nine loci in GENCODE. The *ZSCAN9* protein-coding gene contains eight transcripts in GENCODE v16, four of which are omitted here for clarity. 5′ UTR variation has also been omitted in the transcripts shown, and exon sizes are not to scale. Annotated CDSs are highlighted in green, UTRs in red. Transcript *A* appears to be the major spliceform of the locus based on transcriptomics data (not shown). The putative CDS annotation of transcript *B* was prompted by the identification of polyadenylation features marking a genuine transcript end point (aaaa). Transcripts *C* and *D* contain a cassette exon sharing the same splice donor site (3′ edge), although with differing splice acceptor sites (5′ edge). The larger form of this exon in transcript *D* contains a LoF variant identified by MacArthur and colleagues within 1000 Genomes Project pilot phase data (filled triangle) (MacArthur et al. 2012); the variant is a C/T change that creates an in-frame premature termination codon (PTC).

2013). For these reasons, it is difficult to speculate on what proportion of this experimentally detected transcription will be converted into informative gene annotation.

How can we be sure that an existing transcript model is complete in terms of its exonic structure? Incomplete models are likely to be common in GENCODE, where models can be constructed based on single ESTs. Even mRNAs and cDNAs may not extend to the precise 5′ or 3′ ends of the transcript captured. It can be difficult to predict the functionality of incomplete transcript models. For example, it has been reported that numerous GENCODE lncRNAs may actually represent 3′ UTR sequence from protein-coding genes (Miura et al. 2013), while a putative CDS could be recharacterized as an NMD candidate if the model is extended at the 3′ end. For these reasons, protocols have been devised that capture the true ends of transcripts, i.e., the transcription start site (TSS) and polyadenylation (polyA) site. PolyA-seq harnesses RNA-seq technology to generate short sequence reads from the 3′ ends of RNA molecules (Derti et al. 2012), whereas cap analysis of gene expression (CAGE) tags are reads obtained from the 5′ ends of capped transcripts. While polyA-seq is a novel technique, CAGE has been around for a decade (Shiraki et al. 2003). However, its power has increased with the advent of next generation sequencing (Takahashi et al. 2012), and the protocol has been used as part of both the ENCODE Project (Djebali et al. 2012a) and especially the FANTOM Consortium's efforts to characterize mammalian transcriptomes (The FANTOM Consortium et al. 2005; Suzuki et al. 2009; Kawaji et al. 2011). In both cases, sequencing reads are converted into clusters that are mapped onto the genome, therefore indicating the location of transcript start and endpoints. Where CAGE and polyA-seq clusters correspond to the start or endpoint of existing models respectively, this indicates that the model is complete, allowing functional annotation to proceed with confidence. In fact, these data suggest that large numbers of GENCODE transcript models are not yet full-length; only 35% of model start points were seen to overlap with ENCODE CAGE clusters when these data were compared against GENCODE v7, while 38% of protein-coding genes lacked poly(A) features at this time (Harrow et al. 2012).

Along with RNA-seq, polyA-seq and CAGE will also prove highly useful in the identification of entirely new transcripts, both within existing genes and entirely novel transcribed loci. This process is illustrated in Figure 4, where a series of non-spliced cDNA and EST transcripts are seen to align to an intergenic region of chromosome 6. This transcribed region correlates with the location of CAGE clusters derived from ENCODE data (Djebali et al. 2012a) and polyA-seq clusters generated by Derti et al. (2012). Furthermore, we can integrate RNA-seq data produced by the Illumina Human BodyMap 2.0 project, which captures transcription in 16 human tissues (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513). The read coverage graphs generated from ovary and prostate data by Ensembl (Flicek et al. 2013) are seen to correlate with the region defined between the CAGE and polyA-seq clusters. These three data sets thus combine to identify a novel lncRNA locus with confidence. A limitation of this approach becomes apparent when considering complex AS loci, where the same CAGE and polyA-seq clusters can be linked to several transcripts (and recall also that not all interganic transcripts are polyadenylated). In the near future, however, it seems likely that third generation sequencing platforms will allow us to capture full-length RNA molecules as part of a single protocol. This approach may be particularly powerful when combined with RT-PCR, allowing for the targeted validation of novel transcripts as well as the extension of existing incomplete models (Howald et al. 2012).

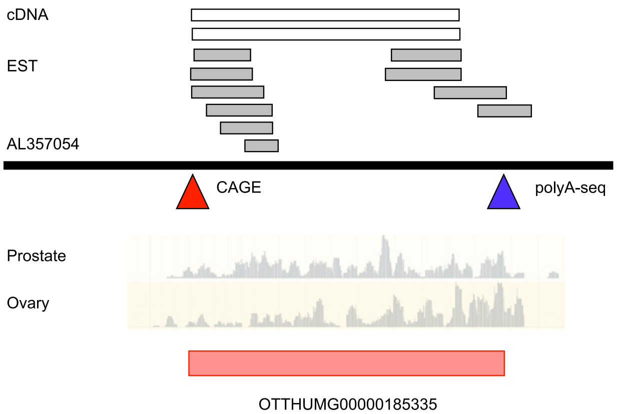

Figure 4. The annotation of a novel lncRNA locus in GENCODE. A schematic diagram of an ∼4-kb region of human chromosome 6 is shown, within BAC AL357054. The alignment of cDNA (accession numbers AX747750, AK092822) and EST (DA816101, DA923061, DA427401, BG541155, AI831721, AW135930, CB052137, DB332016, CB052136, AA004346) data indicate the location of a transcribed locus. Further support comes from the mapping of CAGE and polyA-seq clusters, indicated by red and blue triangles, respectively. CAGE data are taken from the ENCODE project (Djebali et al. 2012a) and polyA-seq data are taken from Derti et al. (2012). For CAGE, co-locating clusters are found in the majority of cells investigated by ENCODE, including primary non-immortalized lines (not shown). Co-locating polyA-seq clusters are derived from brain, testes, and muscle tissues (not shown). *Underneath* is Illumina Human BodyMap 2.0 RNA-seq data from two representative tissues out of 16 available—prostate and ovary—in the form of read coverage graphs. These data were mapped to the genome by Ensembl, using the BWA methodology (Li and Durbin 2009; Flicek et al. 2013). The correspondence between each of these data sets allowed for a new 2814-bp lncRNA model to be built (red rectangle; subcategorized as a lincRNA biotype; see Fig. 1B), accession number OTTHUMG00000185335.

The annotation of CDSs based on proteomics data

Following the capture of the human transcriptome, our focus turns to its functional annotation. An obvious first question to ask of a transcript model is whether it is translated into a protein or peptide molecule. While we may be approaching a final value for the number of human protein-coding genes, the number of functional protein isoforms generated by AS remains hard to estimate. Alternative CDSs can be generated by exon skipping, splice site shifts, or from the use of alternative first or last coding exons. However, we cannot assume that AS within protein-coding genes leads to the translation of stable protein molecules; such transcripts could be noise or perhaps functional noncoding transcripts (Mudge et al. 2011; Ezkurdia et al. 2012; Frankish et al. 2012). Unfortunately, it is still not possible to sequence protein molecules in a manner analogous to RNA or DNA sequencing. However, improvements have been made to mass spectrometry (MS) in recent years, such that "proteogenomics" may now be considered an important tool in genome analysis (Domon and Aebersold 2006; Yates 2013). In particular, state of the art tandem MS can identify peptide sequences with high sensitivity and specificity in a manner approaching high throughput. These data can be publicly accessed via the online databases PRIDE (Vizcaino et al. 2010) and PeptideAtlas (Deutsch 2010), which currently contain hundreds of millions of spectra. Recently, Ezkurida and colleagues were able to identify peptides linked to 35% of the protein-coding genes in

GENCODE (Ezkurdia et al. 2012). Even so, there are limitations to this technique. First, spectra commonly represent short peptides that are unable to distinguish between AS CDSs. Second, both the processing of the spectra and the subsequent genomic mapping are complex, computationally intensive techniques. In particular, peptide to genome mappings may suffer from a high false positive rate unless rigorous methods are used (a situation confounded by the occurrence of AS) (Tanner et al. 2007; Brosch et al. 2011; Ezkurdia et al. 2012).

Ribosome profiling (RP) is a newer technique designed to infer translated regions of RNA molecules (Ingolia et al. 2009). Essentially, the short portion of an RNA that is bound to a ribosome survives a round of chemical degradation and is then recovered and sequenced. Typically, ribosome stalling is induced by drug treatment, allowing for the capture of translation initiation sites (TIS) (Ingolia et al. 2011; Lee et al. 2012). RP thus avoids the complications of dealing with protein molecules directly, and is extremely high-throughput. Recently, Ingolia and colleagues identified thousands of translated regions within the transcriptome of mouse embryonic cells (Ingolia et al. 2011). Of particular interest is their identification of large numbers of TIS found within previously annotated CDSs, including non-ATG codons. Further work in human provides similar findings generally, indicating that RP will be highly useful for both TIS validation and discovery (Lee et al. 2012; Michel et al. 2012). RP also confirms that certain ORFs can be translated from the same transcript as the recognized CDS (Ingolia et al. 2011; Lee et al. 2012; Michel et al. 2012). Figure 5 details the integration of human RP data from Lee et al. (2012) with the GENCODE annotation of the *PNRC2* protein-coding locus. Previously, a 139aa CDS had been annotated in three transcript models that differ in their 5′ UTR confirmations. However, our analysis of RP data supports the usage of a TIS at the start of a previously unrecognized 56aa upstream ORF (uORF). It remains unclear what proportion of such uORFs actually encodes mature polypeptides. An alternative possibility is that they are regulatory in nature, perhaps controlling overall protein output from the locus by sequestering the ribosome and limiting translation from the downstream TIS (Somers et al. 2013). For example, the protein output of *ELK1* is controlled in part via the differential splicing of a uORF in the 5′ UTR (Rahim et al. 2012). Certainly, uORFs represent a further challenge to our preconceptions of transcript functionality, and they are a new focus for transcript annotation.

Comparative genomics can support CDS annotation

While MS and RP analyses look set to become standard techniques for CDS annotation in the near future, we require other methods to examine translation at the present time. Comparative genomics is based on the idea that conservation indicates functionality (Boffelli et al. 2004; Dermitzakis et al. 2005). Of the 20,000 protein-coding genes known in human and mouse, at least 80% can be defined as orthologs (Mouse Genome Sequencing Consortium et al. 2002). Conservation is not only a gene-level attribute, however; it can also be used to describe the individual transcripts and CDS found within a gene. In the former case, we can examine the conservation of TSS, polyadenylation signals, and splice sites, while in the latter case we can compare TIS and termination codons alongside overall amino acid composition. The *PNRC2* locus is a fascinating example in this regard, since we observe that both the upstream and downstream reading frames annotated in Figure 5 are widely conserved amongst vertebrates, including mouse and zebrafish. The conservation argument therefore supports the functionality of both translations, and suggests that this locus may be genuinely bicistronic.

Conservation can also be a powerful technique for inferring the functionality of AS events. We have estimated that approximately one third of human and mouse orthologous gene pairs contain more than one conserved CDS (Mudge et al. 2011). For example, the human *BSG* gene contains four distinct CDSs, linked to alternative first exons and exon-skipping events (Fig. 6). When compared against the annotation of the mouse locus, we observe that two of these CDSs are supported by transcriptional evidence in both species (Pairs A/a and B/b). In other words, conservation indicates that the skipping of the 348-bp cassette exon of model A

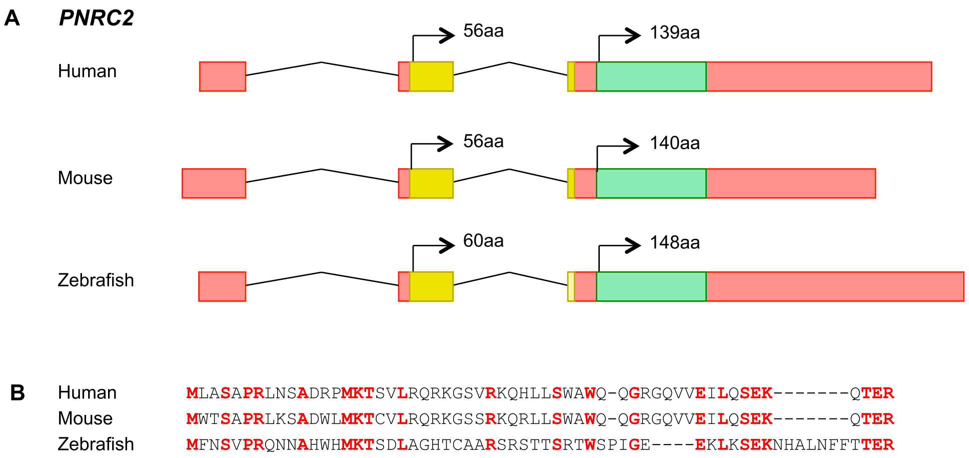
A
PNRC2
Human
56aa
139aa
Mouse
56aa
140aa
Zebrafish
60aa
148aa
B
Human MLASAPRLNSADRPMKTSVLRQRKGSVRKQHLLSWAWQ-QGRGQVVEILQSEK-------QTER
Mouse MWTSAPRLKSADWLMKTCVLRQRKGSSRKQRLLSWAWQ-QGRGQVVEILQSEK-------QTER
Zebrafish MFNSVPRQNNAHWHMKTSDLAGHTCAARSRSTTSRTWSPIGE----EKLKSEKNHALNFFTTER

Figure 5. Annotation of the proline-rich nuclear receptor coactivator 2 locus with ribosome profiling data. (*A*) This schematic shows the GENCODE annotation for the human *PNRC2* locus compared against its mouse and zebrafish orthologs. Annotation for the latter two models is taken from the Vega manual annotation resources (Wilming et al. 2008). Alternative splicing within the 5′ UTR has been omitted for clarity, and intron sizes are not to scale. Equivalent ORFs are shown in yellow and green in each model, with UTR sequences shown in red. The splice donor site of the 5′ UTR exon is conserved between human and mouse; whether this is also true for the 5′ UTR of the zebrafish model cannot be ascertained. The downstream CDS encodes the known PNRC2 protein. The TIS of the upstream ORF is supported by ribosome profiling data in human and mouse, from Lee et al. (2012) and Ingolia et al. (2011), respectively. Data from the latter set also support the TIS of the *PNRC2* CDS in mouse. Equivalent RP resources for zebrafish are not available. (*B*) Alignment of the human, mouse, and zebrafish upstream ORFs, with conserved residues highlighted in red.

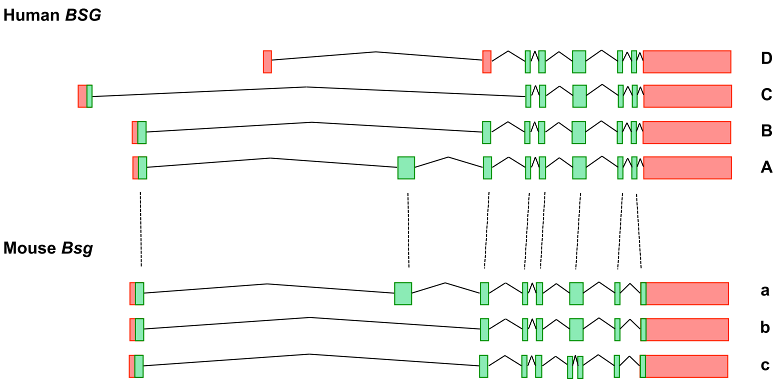

Figure 6. A comparison of alternative splicing within the human and mouse basigin locus. In total the human *BSG* protein-coding locus contains 13 transcript models in GENCODE v16, four of which are shown here for clarity (*A*–*D*). Similarly, three CDSs from the orthologous mouse *Bsg* locus are shown (*a*–*c*), taken from the manual annotation resources of the Vega project (Wilming et al. 2008). Annotated CDSs are highlighted in green, UTRs in red. The human locus contains an intron 5 bp downstream from the termination codon that has no counterpart in the mouse locus; other minor UTR variations have been omitted. Dashed lines indicate exon level orthology; intron sizes are to approximate scale only. Transcript models *A* and *a*, and *B* and *b*, are thus equivalent between human and mouse, and each is supported by transcriptional evidence in both species.

is a functional AS event. Interestingly, conservation can also indicate the functionality of NMD-linked transcription (see Toward a Definition of Transcript Functionality section). The NMD process is commonly linked to the inclusion of a "poison exon" in a transcript, defined as an exon that introduces a PTC into the CDS. It is apparent that such exons can show strong conservation, even across the vertebrate clade; we estimate that 10% of human protein-coding genes contain NMD transcripts that are conserved in mouse (Mudge et al. 2011).

The most obvious limitation of the conservation argument is that it cannot judge, for example, the functionality of the thousands of human and mouse AS CDSs that are not conserved between the two species. The cassette exon of human *ZSCAN9* model D that contains a LoF mutation has no mouse counterpart (Fig. 3), and neither does the alternative first exon of human *BSG* model C (Fig. 6). Similarly, conservation cannot be used to judge species-specific loci. Lineage specific protein-coding loci are commonly referred to as "orphans." Their true prevalence in the human genome remains a source of uncertainty (Khalturin et al. 2009; Brosch et al. 2011; Tautz and Domazet-Loso 2011; Neme and Tautz 2013), particularly given that genuine CDSs can be under 300 bp in length, and yet random ORFs of a similar size are commonly found on transcripts due to chance (Clamp et al. 2007). The functionality of lineage specific CDSs must therefore be judged through the integration of proteomics data. A further limitation of the conservation proxy is that the correct in silico identification of orthologous exons becomes more difficult as the genomes targeted become more diverged. In our experience, even genuine exonic alignments between human and mouse are commonly missed by computational analysis, in particular when considering AS exons that are short (Mudge et al. 2011). Also, if conservation is measured as constraint on the genome sequence, the proxy can highlight pseudogenes as functional loci. Finally, the usefulness of comparative annotation depends entirely upon the availability of high-quality genome sequences and, ideally, large pools of transcriptomic data. In practice, few vertebrate species are comparable with human and mouse in this regard.

Toward a functional annotation of lncRNAs

Unlike for CDS transcripts, where a strong paradigm for their functionality existed prior to the genome-sequencing era, efforts to understand the functional role of lncRNAs are proceeding in parallel with their discovery. Unfortunately, while CDS transcripts and small RNAs can be readily identified by ab initio methods, there are no known sequence motifs or secondary structures that appear common amongst lncRNAs (Gorodkin and Hofacker 2011). This complicates efforts to annotate lncRNA data sets in a meaningful way. In fact, the true extent of functionality in this transcript category has been a source of debate since its identification. Our knowledge of lncRNAs evolved from observations of pervasive transcription across eukaryotic genomes, which was originally suggested using genomic tiling arrays (Kapranov et al. 2002; Rinn et al. 2003). However, this technology is known to be prone to experimental artifacts that could lead to false hybridization signals (Johnson et al. 2005). In due course, the presence of widespread noncoding transcription was also supported by sets of mammalian noncoding cDNAs (The FANTOM Consortium et al. 2005) and improved array-based analyses (The ENCODE Project Consortium 2007; Kapranov et al. 2007). Concerns regarding the biological relevance of such transcripts resurfaced, however, following initial observations that they show low levels of sequence conservation (Wang et al. 2004). In fact, this highlights a further issue with the use of constraint as a proxy for functionality: The best way to measure constraint remains debatable, and different approaches can yield quite different results. Of particular significance is the method by which the neural rate of evolution is estimated, since this is essentially the "background reading" against which constraint is measured. Pheasant and Mattick (2007) and Ponting and Hardison (2011) have provided detailed discussions on this issue. Significant differences can also arise from the nature of the sequence alignments performed. Specifically, whole-genome comparisons are typically based around "windows" of alignment, and these may lack the granularity needed to uncover short, dispersed regions of conservation such as those that have been identified between the human and mouse *HOTAIR* orthologs (Pang et al. 2006; Schorderet and Duboule 2011).

In our view, it is not yet clear how useful constraint will be in the functional annotation of lncRNAs. One the one hand, Ponjavic and colleagues have shown that purifying selection within lncRNA exons is more readily detected when robust analytical methods are used, in particular when focusing on specific regions of the locus such as the promoter region and splice sites (Ponjavic et al. 2007). On the other, functional lncRNAs such as *HOTAIR* may contain a bulk of sequence that does not contribute to their actual function and so does not experience constraint (Tsai et al. 2010). Indeed, lncRNAs that represent by-products of functional transcription—such as *Airn* (Latos et al. 2012)—may not experience purifying selection at all, given that their sequence content may be largely unimportant. In fact, it is apparent that the majority of lncRNA transcripts are subjected to rapid evolutionary turnover in the mammalian order (Cabili et al. 2011; Kutter et al. 2012). This may imply that functionality across lncRNA data sets is rapidly evolving, and a link has been postulated with the significant changes in protein gene expression levels witnessed between mammalian species (Kutter et al. 2012). At the present time, a role in the regulation of gene expression looks set to become a central paradigm of lncRNA functionality (Mattick 2007; St. Laurent and Wahlestedt 2007; Guttman and Rinn 2012; Moran et al. 2012; Rinn and Chang 2012). This has in turn led to suggestions that it may be possible to identify functional lncRNAs by capturing transcripts that interact with chromatin-modification complexes. For example, Zhao and colleagues used high-throughput RNA immunoprecipitation to identify several hundred mouse lncRNAs that bind Polycomb repressive complex 2 (Zhao et al. 2010). Nonetheless, it is currently uncertain as to exactly how these interactions should be interpreted in terms of functionality (Rinn and Chang 2012; Brockdorff 2013).

Clearly, we still have much to learn about how lncRNAs function, and it may not in fact be appropriate to regard this as a single homogenous class of transcript. Of particular note here is the recent discovery by St. Laurent and colleagues that 10% of our genome may consist of "very long intergenic ncRNAs" (vlincRNAs) (St. Laurent et al. 2013). Such transcripts, which are over 50 kb in size, have been previously shown to contribute the bulk of non-ribosome-associated, non-mitochondrial RNA in some human cells (Kapranov et al. 2010). However, the relationship between proposed vlincRNAs and the shorter lncRNAs that exist in GENCODE is currently unclear. For such reasons, it is difficult to speculate on the proportion of the 22,444 lncRNA transcripts annotated in GENCODE that have genuine functionality at the present time. Instead, the value of annotation projects to the description of lncRNAs begins with the observation that large numbers of lncRNA models are likely to be either incomplete or entirely missing from our gene sets. We believe that our initial focus should therefore be on the generation of a comprehensive set of complete transcript structures, onto which biological information can be layered as it becomes available. Certainly, it is true that those few lncRNA genes that are well understood in terms of function—such as *HOTAIR* and *Airn*—are so because they have been dissected in detailed laboratory studies. On the other hand, the existence of putatively annotated lncRNA models will likely prove an invaluable resource in the design of experiments to study individual loci in detail.

The annotation of gene expression data

Gene expression data may provide further insights into the functionality of both protein-coding and lncRNA transcripts. We can start with the following logic: If a transcript is abundant in the cell, this suggests it may be functional. This is based on a presumption that, while stochastic noise may be common generally, specific "errors" in transcription or splicing are rare. This logic is implicit in the annotation of the transcript model in Figure 3, where appreciable read coverage is observed in several tissues. As RNA-seq technologies improve, it should become routine to measure the expression of every transcript in a gene set, across a wide range of tissues and developmental stages. This will also highlight models that have consistently very low or irreproducible levels of expression. Furthermore, a potential indication of functionality can be also provided by the demonstration of restricted expression, i.e., where a transcript displays tissue or developmental specificity. By this argument, restricted expression suggests that the transcript is generated via gene regulation as opposed to stochastic noise, although care must be taken to ensure that the observation is not simply due to false negative detection in certain experiments linked to low expression levels. This logic is frequently incorporated into studies examining AS within mammalian genes based on RNA-seq (Pan et al. 2008; Wang et al. 2008), and it appears that >50% of AS events in human are tissue specific (Wang et al. 2008). Figure 7 illustrates the use of expression profiling in AS annotation, focusing on the *ACSL6* gene. This gene contains a pair of adjacent 78-bp cassette exons that are homologous and spliced in a mutually exclusive manner (based on EST data). On integrating Illumina Human BodyMap RNA-seq data, we see evidence of tissue specificity in this AS: The upstream exon is transcribed in lymph, lung, and adrenal cells while the downstream is not, whereas the opposite pattern is observed in liver cells. In contrast, both exons appear to be utilized in brain tissue. Restricted expression has also been used to infer functionality within lncRNA sets, and it appears that tissue specificity is significantly more common amongst lncRNA transcripts compared with protein-coding transcripts (Cabili et al. 2011; Derrien et al. 2012; Hangauer et al. 2013). In addition, St. Laurent and colleagues have recently demonstrated the potential value of physiological time-course experiments in identifying large numbers of novel tran-

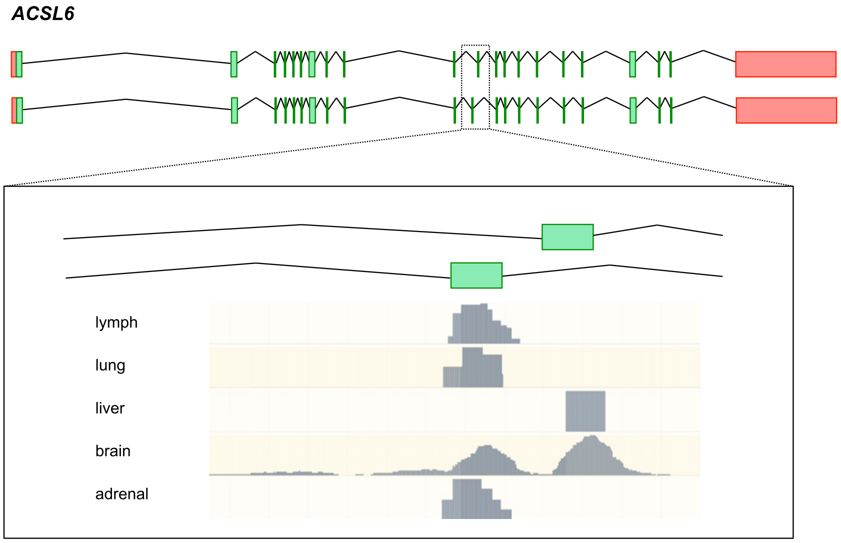

Figure 7. Tissue-specific alternative splicing within the human acyl-CoA synthetase long-chain family member 6 protein-coding gene. The GENCODE v16 annotation of *ACSL6* contains 21 AS transcripts, two of which are shown here. The transcripts differ in the incorporation of distinct members of a mutually exclusive 78-bp cassette exon pair found in the same intron. Illumina Human BodyMap RNA-seq data read graphs are shown for five tissues (produced by Ensembl; see Fig. 4 legend for further details). Lymph, lung, and adrenal cells are seen to apparently utilize the upstream exon only, whereas liver cells show the opposite pattern. Brain cells appear to utilize either exon. Though apparently homologous, the exons differ at 30 out of 78 bp and 10 out of 26aa (not shown), which provides confidence that the read mapping has not been confounded by paralogy.

scripts whose expression level changes significantly in response to changing cellular conditions (St. Laurent et al. 2012).

Expression profiling will also benefit laboratory studies. For example, such data can show us that to study a particular protein we may need to focus on a specific tissue. It could be argued that we will not have a fully annotated transcriptome until we have captured the expression profile of every transcript. Following this logic, Gonzalez-Porta and colleagues have recently compared the expression levels of the different alternative transcripts located within GENCODE protein-coding genes using RNA-seq (Gonzalez-Porta et al. 2013). The authors found that 85% of cellular mRNA represents the combined expression of a single "major transcript" from each locus. However, when differentiating among the 16 Illumina Human BodyMap tissues, transcription was seen to "switch" to an alternative transcript in 35% of genes. Unexpectedly, they also found that the major transcript of 20% of genes is not annotated with a CDS (typically as a retained intron model), indicating that the link between transcript abundance and functionality may not always be straightforward. This illustrates an important factor in expression profiling: Researchers have to decide how to distinguish significant transcription from background noise. ENCODE judged significance based on the irreproducible discovery rate (Li et al. 2011); others have used simple read-depth thresholds such as reads per kilobase per million mapped reads (Mortazavi et al. 2008) or fragments per kilobase of exon per million fragments mapped (Trapnell et al. 2010). However, we caution that transcripts with very low expression levels may be functional (Clark et al. 2011), while transcripts with strong expression profiles may have undergone recent pseudogenization, losing their function although not their transcription potential.

Notably, the expression levels of lncRNAs are on average at least 10 times lower than for protein-coding transcripts (Derrien et al. 2011). There are additional caveats for interpreting the expression profiles of lncRNAs, relating to unanswered questions about their function and evolution. First, there are well-established scenarios where nonfunctional lncRNAs could display restricted expression. Most obviously, transcripts found exclusively in cancer-derived libraries should concern researchers, since cancer cells have a well-established tendency for aberrant transcription (Ghigna et al. 2008). Furthermore, there is evidence that both embryonic stem cells and testis cells undergo "hypertranscription," whereby even tissue-specific genes become expressed at detectable levels (Efroni et al. 2008; Kaessmann 2010). In both cases this phenomenon has been linked to the presence of constitutively open chromatin. In fact, the abundance of novel transcripts detected in testis (especially lncRNAs [Cabili et al. 2011]) has led to the suggestion that this organ may represent a breeding ground for new genes; this may be due to the particularly efficient activity of proto-promoters in testis cells (Kleene 2005; see below). Finally, as discussed in the following section, many lncRNAs are associated with either enhancer elements or the bidirectional promoters of protein-coding genes. If such genomic sequences operate in a tissue- or developmentally specific manner, then the associated lncRNA transcripts could theoretically display the same pattern of expression even if they are nonfunctional. As such, the true value of expression profiling in inferring the functionality of such transcripts will not become clear until we find out more about their biological nature. Conversely, the expression profiling of annotated lncRNAs is likely to prove of great assistance in this regard. The fact that the novel transcript identified in Figure 3 shows strong expression in prostate and ovary tissues may not confirm its functionality, although it does provide a valuable starting point for scientists who wish to study this locus in more detail.

Can regulatory signals in the genome aid transcript annotation?

Gene expression is controlled by sequences encoded in the genome such as promoter regions (Carninci et al. 2006), enhancer elements (Kulaeva et al. 2012), and splicing signals (Barash et al. 2010), and also by epigenomic signatures such as chromatin modifications (Hoffman et al. 2013) and DNA methylation (Smith and Meissner 2013). Can the description of such elements be used to aid functional transcript annotation? It seems plausible that the annotation of AS could benefit from the description of regulatory signals. For years, it has been clear that AS is at least partially directed by sequences found in the exons or introns of the nascent RNA, and a variety of splicing enhancer and silencer elements are now known (Chen and Manley 2009). By focusing on such motifs, Barash and colleagues combined transcriptomics and machine learning to show that it is possible to predict the patterns of splice site usage in genes given only the genome sequence (Barash et al. 2010). While the control of AS is a highly complex process that remains improperly understood, this work suggests it may be possible to fully decipher a "splicing code" in the future. If so, the boon to annotation would be significant (Irimia and Blencowe 2012); a comprehensive splicing code could provide confidence that annotated AS events are non-spurious, and also predict functional AS transcripts that have not yet been captured by sequencing projects.

The value of epigenomics and promoter mapping to gene annotation is harder to gauge. ENCODE and other projects have dedicated significant resources to the functional annotation of promoters, other *cis*-regulatory regions, and epigenomic marks on the human genome (Barski et al. 2007; The ENCODE Project Consortium 2007; Mikkelsen et al. 2007; Rando and Chang 2009; The ENCODE Project Consortium 2012; Neph et al. 2012; Sanyal et al. 2012; Thurman et al. 2012). Extensive data are available, for example, on the mapping of a wide variety of histone modification patterns as well as the sites of occupancy of certain transcription factors; both can be indicative of promoter elements. All eukaryotic primary transcripts are theoretically generated by promoters, which suggests that promoter mapping could be used to identify novel transcripts. An obvious caveat is that promoter mapping is essentially a locus level technique; it will not distinguish between AS transcripts that share similar TSS. These data may instead be of more use in the identification of novel lncRNAs. In this manner, Guttman and colleagues identified 1600 multiexon mouse lncRNAs, using the histone modifications H3K4me3 and H3K36me3 as markers for promoter sequences and transcribed regions respectively, in combination with DNA tiling arrays (Guttman et al. 2009). In a follow up study, the authors found that significant numbers of their lncRNAs interact directly with chromatin regulatory proteins in ES cells, suggesting that this proxy has identified functional loci (Guttman et al. 2011).

However, we can also envisage situations where an association between a promoter region and a lncRNA locus may not confirm functionality. In particular, eukaryotic promoters are frequently bidirectional (Carninci et al. 2006; Cabili et al. 2011), such that lncRNAs are commonly found antisense to protein-coding gene promoters (Fig. 1B, model ix; Core et al. 2008; Sigova et al. 2013). This transcription could theoretically be a by-product of chromatin remodeling, although certain antisense transcripts have been shown to regulate transcription of the neighboring gene (Kanhere et al. 2010). Furthermore, noncoding transcription also commonly co-localizes with enhancer elements, and the potential functionality of this process remains similarly ambiguous (Kim et al. 2010). LncRNAs in GENCODE are also commonly linked to pseudogenes of all categories. Conceivably, active promoters may persist at genes that have been pseudogenized, and could generate either noisy transcription as they move toward an inactive state or lncRNAs with novel functionality. Processed pseudogenes may also be transcribed, and evidence suggests that regulatory regions close to the insertion site frequently drive this transcription (Vinckenbosch et al. 2006; Kaessmann et al. 2009). While such transcription may commonly be opportunistic, retrotransposition should nonetheless be considered an additional mechanism for the "birth" of new lncRNAs (Kaessmann 2010).

Second, the annotation potential of epigenomics is hampered by gaps in our understanding of promoter sequences. If a nascent lncRNA locus does not co-opt an existing promoter, it is presumably transcribed from a novel promoter. It is thought that the core eukaryotic promoter can arise by de novo mutation (Kaessmann 2010); alternately, the FANTOM project has identified thousands of human and mouse CAGE clusters found within TEs (Faulkner et al. 2009). One can then imagine a scenario where a "proto-promoter" gives birth to a novel, nonfunctional lncRNA. Over time, the lncRNA may develop functionality, and in parallel the proto-promoter may become more elaborate as it picks up sequence motifs that confer gene regulation. The co-localization of a complex promoter with a lncRNA would then suggest the functionality of that locus. While a "community standard" definition of a complex promoter is not yet available, work is now underway to classify human promoters based on the genome sequences and epigenetic marks found in association. For example, ENCODE have pooled their data set of identified chromatin elements to generate a "segmentation" analysis of promoter regions, and find a clear correlation between the marks these data produce and the annotated 5′ ends of GENCODE protein-coding loci (Hoffman et al. 2013). On the other hand, a minority of GENCODE lncRNA are currently linked to both promoter-associated chromatin modifications and TF binding sites (Djebali et al. 2012a). It may be that lncRNA promoters are generally less elaborate than those of protein-coding genes, and this may be because these loci do not need to respond to such a wide variety of gene expression factors in order to perform their function. However, this scenario is also consistent with the existence of widespread noise amongst lncRNA gene sets. Alternatively, this lack of association may also be due to the structural incompleteness typical of lncRNA transcript models.

Summary

We believe the question "what is a gene?" conceals a question of more pressing importance: Which transcripts are functional, and how do they function? There are two approaches we can take to tackle this question. First, the true confirmation of transcript functionality, and a detailed understanding of the nature of this functionality, can only be gained in the laboratory. Nonetheless, the number of identified human transcripts likely exceeds 200,000, the significant majority of which have to be examined by in-depth single gene studies. Modern genomics is therefore going through an awkward transition period: We know that transcriptional complexity exists, yet our understanding of the functional basis of this complexity remains imperfect. Certainly, nonfunctional transcripts do occur, and they can confound our scientific analyses. Consider the scientist tasked with judging the effects of several hundred variant sites: First, she will want to know which of these overlap transcripts; second, if and how these transcripts actually function. Modern genomics (and indeed medicine) demands to understand the entirety of the genome and transcriptome right now, and to match this demand we have to turn to the second approach, which is to predict functionality by combining next-generation data sets with computational analyses. It is important to understand the limitations to what this approach can achieve. However, we believe that many of the caveats described above are temporary hurdles. In particular, a significant step forward will be taken when short-read technologies are replaced by techniques able to sequence entire transcripts, while comparative annotation will gain power from the availability of more high-quality species genomes and transcriptomes. To be clear, the purpose of this strategy is not to replace single gene studies, which will always be a vital part of science. Instead these strategies should be seen as complimentary; functional annotation is greatly improved by scientific advances made in the laboratory, while targeted gene studies can benefit enormously by considering predictions made by annotation. Finally, no one knows what proportion of the transcriptome is functional at the present time; therefore, the appropriate scientific position to take is to be open-minded. We thus do not claim that the annotation of the human genome is close to completion. If anything, it seems as if the hard work is just beginning.

Acknowledgments

We thank Alvis Brazma for informative discussion on the issues covered in this manuscript. This work was supported by the National Institutes of Health (grant no. 5U54HG004555) and the Wellcome Trust (grant no. WT098051).

References

Ball P. 2013. DNA: Celebrate the unknowns. *Nature* **496:** 419–420.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465:** 53–59.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* **5:** 456–465.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348.

Brinkman BM. 2004. Splice variants as cancer biomarkers. *Clin Biochem* **37:** 584–594.

Brockdorff N. 2013. Noncoding RNA and Polycomb recruitment. *RNA* **19:** 429–442.

Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, Verstraten R, Adams DJ, Harrow J, Choudhary JS, et al. 2011. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res* **21:** 756–767.

Brosius J. 2005. Waste not, want not—transcript excess in multicellular eukaryotes. *Trends Genet* **21:** 287–288.

Brosius J. 2009. The fragmented gene. *Ann NY Acad Sci* **1178:** 186–193.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25:** 1915–1927.

Camacho-Vanegas O, Camacho SC, Till J, Miranda-Lorenzo I, Terzo E, Ramirez MC, Schramm V, Cordovano G, Watts G, Mehta S, et al. 2012. Primate genome gain and loss: A bone dysplasia, muscular dystrophy, and bone cancer syndrome resulting from mutated retroviral-derived MTAP transcripts. *Am J Hum Genet* **90:** 614–627.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626–635.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10:** 741–754.

Chen L, Tovar-Corona JM, Urrutia AO. 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Hum Mol Genet* **20:** 4422–4429.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci* **104:** 19428–19433.

Clark MB, Mattick JS. 2011. Long noncoding RNAs in cell biology. *Semin Cell Dev Biol* **22:** 366–376.

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A et al. 2011. The reality of pervasive transcription. *PLoS Biol* **9:** e1000625.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.

Crick F. 1970. Central dogma of molecular biology. *Nature* **227:** 561–563.

Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences— an unexpected feature of mammalian genomes. *Nat Rev Genet* **6:** 151–157.

Derrien T, Guigo R, Johnson R. 2011. The long non-coding RNAs: A new (p)layer in the "dark matter." *Front Genet* **2:** 107.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22:** 1775–1789.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22:** 1173–1183.

Deutsch EW. 2010. The PeptideAtlas Project. *Methods Mol Biol* **604:** 285–296.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012a. Landscape of transcription in human cells. *Nature* **489:** 101–108.

Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, Howald C, Foissac S, Ucla C, Chrast J, et al. 2012b. Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS ONE* **7:** e28213.

Domon B, Aebersold R. 2006. Mass spectrometry and protein analysis. *Science* **312:** 212–217.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci* **110:** 5294–5300.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, et al. 2012. The UCSC Genome Browser database: Extensions and updates 2011. *Nucleic Acids Res* **40:** D918–D923.

Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, Bazett-Jones DP, Le Grice S, McKay RD, Buetow KH, et al. 2008. Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell* **2:** 437–447.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Ezkurdia I, Del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, Valencia A, Tress ML. 2012. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol Biol Evol* **29:** 2265–2283.

The FANTOM Consortium, Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41:** 563–571.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55.

Frankish A, Mudge JM, Thomas M, Harrow J. 2012. The importance of identifying alternative splicing in vertebrate genome annotation. *Database (Oxford)* **2012:** bas014.

Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, Del Pozo A, Tress M, Johnson R, Guigo R, et al. 2012. Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res* **22:** 1231–1242.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17:** 669–681.

Ghigna C, Valacca C, Biamonti G. 2008. Alternative splicing and tumor progression. *Curr Genomics* **9:** 556–570.

Gingeras TR. 2007. Origin of phenotypes: Genes and transcripts. *Genome Res* **17:** 682–690.

Gingeras TR. 2009. Implications of chimaeric non-co-linear transcripts. *Nature* **461:** 206–211.

Gonzalez-Porta M, Calvo M, Sammeth M, Guigo R. 2012. Estimation of alternative splicing variability in human populations. *Genome Res* **22:** 528–538.

Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol* **14:** R70.

Gorodkin J, Hofacker IL. 2011. From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comput Biol* **7:** e1002100.

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* **5:** 578–590.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482:** 339–346.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477:** 295–300.

Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9:** e1003569.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774.

Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, Welsh-Bohmer KA, Hulette CM, Denny TN, Goldstein DB. 2008. Tissue-specific genetic control of splicing: Implications for the study of complex traits. *PLoS Biol* **6:** e1.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41:** 827–841.

Howald C, Tanzer A, Chrast J, Kokocinski F, Derrien T, Walters N, Gonzalez JM, Frankish A, Aken BL, Hourlier T, et al. 2012. Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res* **22:** 1698–1710.

Hsu SN, Hertel KJ. 2009. Spliceosomes walk the line: Splicing errors and their impact on cellular function. *RNA Biol* **6:** 526–530.

Huang L, Lou CH, Chan W, Shum EY, Shao A, Stone E, Karam R, Song HW, Wilkinson MF. 2011. RNA homeostasis governed by cell type-specific and branched feedback loops acting on NMD. *Mol Cell* **43:** 950–961.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324:** 218–223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147:** 789–802.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Irimia M, Blencowe BJ. 2012. Alternative splicing: Decoding an expansive regulatory layer. *Curr Opin Cell Biol* **24:** 323–332.

Johnson JM, Edwards S, Shoemaker D, Schadt EE. 2005. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21:** 93–102.

Johnsson P, Ackley A, Vidarsdottir L, Lui WO, Corcoran M, Grander D, Morris KV. 2013. A pseudogene long-noncoding-RNA network regulates *PTEN* transcription and translation in human cells. *Nat Struct Mol Biol* **20:** 440–446.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20:** 1313–1326.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat Rev Genet* **10:** 19–31.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* **478:** 483–489.

Kanhere A, Viiri K, Araujo CC, Rasaiyaah J, Bouwman RD, Whyte WA, Pereira CF, Brookes E, Walker K, Bell GW, et al. 2010. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell* **38:** 675–688.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316:** 1484–1488.

Kapranov P, St. Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF, et al. 2010. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol* **8:** 149.

Kawaji H, Severin J, Lizio M, Forrest AR, van Nimwegen E, Rehli M, Schroder K, Irvine K, Suzuki H, Carninci P, et al. 2011. Update of the FANTOM web resource: From mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res* **39:** D856–D860.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: Are taxonomically-restricted genes important in evolution? *Trends Genet* **25:** 404–413.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:** 182–187.

Kleene KC. 2005. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. *Dev Biol* **277:** 16–26.

Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. 2013. Gene regulation by the act of long non-coding RNA transcription. *BMC Biol* **11:** 59.

Kozomara A, Griffiths-Jones S. 2011. miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39:** D152–D157.

Kulaeva OI, Nizovtseva EV, Polikanov YS, Ulianov SV, Studitsky VM. 2012. Distant activation of transcription: Mechanisms of enhancer action. *Mol Cell Biol* **32:** 4892–4897.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8:** e1002841.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446:** 926–929.

Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE, et al. 2012. Airn transcriptional overlap, but not its lncRNA products, induces imprinted *Igf2r* silencing. *Science* **338:** 1469–1472.

Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci* **109:** E2424–E2432.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5:** 1752–1779.

Livyatan I, Harikumar A, Nissim-Rafinia M, Duttagupta R, Gingeras TR, Meshorer E. 2013. Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. *Nucleic Acids Res* **41:** 6300–6315.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335:** 823–828.

Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TW, Collins JE, Turner DJ. 2010. FRT-seq: Amplification-free, strand-specific transcriptome sequencing. *Nat Methods* **7:** 130–132.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat Rev Genet* **12:** 671–682.

Mattick JS. 2003. Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25:** 930–939.

Mattick JS. 2007. A new paradigm for developmental biology. *J Exp Biol* **210:** 1526–1547.

Mattick JS. 2011. Long noncoding RNAs in cell and developmental biology. *Semin Cell Dev Biol* **22:** 327.

Melamud E, Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Res* **37:** 4873–4886.

Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC. 2004. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet* **36:** 1073–1078.

Mercer TR, Mattick JS. 2013. Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Res* **23:** 1081–1088.

Michel AM, Roy Choudhury K, Firth AE, Ingolia NT, Atkins JF, Baranov PV. 2012. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* **22:** 2219–2229

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.

Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. 2013. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res* **23:** 812–825.

Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* **34:** 177–180.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464:** 773–777.

Moran VA, Perera RJ, Khalil AM. 2012. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res* **40:** 6391–6400.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J. 2011. The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol Biol Evol* **28:** 2949–2959.

Munsky B, Neuert G, van Oudenaarden A. 2012. Using gene expression noise to understand gene regulation. *Science* **336:** 183–187.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14:** 117.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489:** 83–90.

Ozsolak F, Milos PM. 2011. RNA sequencing: Advances, challenges and opportunities. *Nat Rev Genet* **12:** 87–98.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415.

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends Genet* **22:** 1–5.

Pearson H. 2006. Genetics: What is a gene? *Nature* **441:** 398–401.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. *Genome Biol* **13:** R51.

Pennisi E. 2007. Genomics. DNA study forces rethink of what it means to be a gene. *Science* **316:** 1556–1557.

Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. *Genome Res* **17:** 1245–1253.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010a. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.

Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010b. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6:** e1001236.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17:** 556–565.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res* **21:** 1769–1776.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33:** D501–D504.

Rahim G, Araud T, Jaquier-Gubler P, Curran J. 2012. Alternative splicing within the *elk-1* 5′ untranslated region serves to modulate initiation events downstream of the highly conserved upstream open reading frame 2. *Mol Cell Biol* **32:** 1745–1756.

Rando OJ, Chang HY. 2009. Genome-wide views of chromatin structure. *Annu Rev Biochem* **78:** 245–271.

Raser JM, O'Shea EK. 2005. Noise in gene expression: Origins, consequences, and control. *Science* **309:** 2010–2013.

Raz T, Kapranov P, Lipson D, Letovsky S, Milos PM, Thompson JF. 2011. Protocol dependence of sequencing-based gene expression measurements. *PLoS ONE* **6:** e19287.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81:** 145–166.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, et al. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev* **17:** 529–540.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, et al. 2010. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7:** 909–912.

Sam LT, Lipson D, Raz T, Cao X, Thompson J, Milos PM, Robinson D, Chinnaiyan AM, Kumar-Sinha C, Maher CA. 2011. A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS ONE* **6:** e17305.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489:** 109–113.

Schorderet P, Duboule D. 2011. Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS Genet* **7:** e1002071.

Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, Sato S, Davidson BL, Xing Y. 2011. Widespread establishment and regulatory impact of Alu exons in human genes. *Proc Natl Acad Sci* **108:** 2837–2842.

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* **100:** 15776–15781.

Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, et al. 2013. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci* **110:** 2876–2881.

Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: The vanguard of genome defence. *Nat Rev Mol Cell Biol* **12:** 246–258.

Sleutels F, Zwart R, Barlow DP. 2002. The non-coding air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810–813.

Smith ZD, Meissner A. 2013. DNA methylation: Roles in mammalian development. *Nat Rev Genet* **14:** 204–220.

Somers J, Poyry T, Willis AE. 2013. A perspective on mammalian upstream open reading frame function. *Int J Biochem Cell Biol* **45:** 1690–1700.

Sorek R. 2007. The birth of new exons: Mechanisms and evolutionary consequences. *RNA* **13:** 1603–1608.

St. Laurent G III, Wahlestedt C. 2007. Noncoding RNAs: Couplers of analog and digital information in nervous system function? *Trends Neurosci* **30:** 612–621.

St. Laurent G, Shtokalo D, Tackett MR, Yang Z, Eremina T, Wahlestedt C, Urcuqui-Inchima S, Seilheimer B, McCaffrey TA, Kapranov P. 2012. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics* **13:** 504.

St. Laurent G III, Shtokalo D, Dong B, Tackett MR, Fan X, Lazorthes S, Nicolas E, Sang N, Triche TJ, McCaffrey TA, et al. 2013. VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol* **14:** R73.

Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJ, et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41:** 553–562.

Takahashi H, Kato S, Murata M, Carninci P. 2012. CAGE (cap analysis of gene expression): A protocol for the detection of promoter and transcriptional networks. *Methods Mol Biol* **786:** 181–200.

Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V. 2007. Improving gene annotation using peptide mass spectrometry. *Genome Res* **17:** 231–239.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12:** 692–702.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489:** 75–82.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329:** 689–693.

Uechi T, Tanaka T, Kenmochi N. 2001. A complete map of the human ribosomal protein genes: Assignment of 80 genes to the cytogenetic map and implications for human disorders. *Genomics* **72:** 223–230.

The UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40:** D71–D75.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci* **103:** 3220–3225.

Vizcaino JA, Cote R, Reisinger F, Barsnes H, Foster JM, Rameseder J, Hermjakob H, Martens L. 2010. The Proteomics Identifications database: 2010 update. *Nucleic Acids Res* **38:** D736–D742.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43:** 904–914.

Wang GS, Cooper TA. 2007. Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8:** 749–761.

Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J, Wong GK. 2004. Mouse transcriptome: Neutral evolution of 'non-coding' complementary DNAs. *Nature* doi: 10.1038/nature03016.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.

Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* **36:** D753–D760.

Yates JR III. 2013. The revolution and evolution of shotgun proteomics for large-scale proteome analysis. *J Am Chem Soc* **135:** 1629–1640.

Zarnack K, Konig J, Tajnik M, Martincorena I, Eustermann S, Stevant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152:** 453–466.

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. 2010. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40:** 939–953.

Received May 30, 2013; accepted in revised form September 26, 2013.