# 6 Non-coding RNA characterization

## Many novel and previously known non-coding RNA species are characterized in ENCODE

In addition, we annotated 8,801 automatically derived small RNAs and 9,640 manually curated long non-coding RNA (lncRNA) loci [17]. Comparing lncRNAs to other ENCODE data indicates that lncRNAs are generated through a pathway similar to that for protein coding genes[17]. The GENCODE project also annotated 11,224 pseudogenes, of which 863 were transcribed and associated with active chromatin[18].

We sequenced RNA[16] from different cell lines and multiple subcellular fractions to develop an extensive RNA expression catalogue. Using a conservative threshold to identify regions of RNA activity, 62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons. Of these bases, only 5.5% are explained by GENCODE exons. The majority of transcribed bases are within or overlapping annotated genes boundaries (*i.e.* intronic) and only 31% of bases in sequenced transcripts were intergenic[16].

We used CAGE-seq (5' cap-targeted RNA isolation and sequencing) to identify 62,403 transcription start sites (TSSs) at high confidence (IDR of 0.01) in Tier 1 and 2 cell types. Of these, 27,362 (44%) are within 100 bp of the 5' end of a GENCODE-annotated transcript or previously reported full-length mRNA. The remaining regions predominantly lie across exons and 3' UTRs, and some exhibit cell type restricted expression; these may represent the start sites of novel, cell type-specific transcripts.

Finally, we saw a significant proportion of coding and non-coding transcripts processed into steady state stable RNAs shorter than 200 nucleotides. These precursors include t-, mi-, sn- and sno-RNAs and the 5' termini of these processed products align with the capped 5' end tags[16].

### Detection of annotated and novel transcripts

The Gencode gene (Supplementary Fig. 3a) and transcript (Supplementary Fig. 3b) reference annotation[8] captures our current understanding of the polyadenylated human transcriptome. In the samples interrogated here, we cumulatively detected 70% of annotated splice junctions, transcripts and genes (Fig. 1 and Table 1a). We also detected approximately 85% of annotated exons with an average coverage by RNA-seq contigs of 96%. The variation in the proportion of detected elements among cell lines was small (Fig. 1, width of box plots). Consistent with earlier studies, most annotated elements are present in both polyadenylated (Supplementary Table 3a) and non-polyadenylated (Supplementary Table 3b) samples12-15. Only a small proportion of Gencode elements (0.4% of exons, 2.8% of splice sites, 3.3% of transcripts and 4.7% of genes) are detected exclusively in the non-polyadenylated RNA fraction.

Beyond the Gencode annotated elements, we observed a substantial number of novel elements represented by reproducible RNA-seq contigs. These novel elements covered 78% of the intronic nucleotides and 34% of the intergenic sequences (Supplementary Fig. 4). Overall, the unique contribution of each cell line to the coverage of the genome tends to be small and similar for each cell line (Supplementary Fig. 5). We used the Cufflinks algorithm (see Supplementary Information), and predicted over all long RNA-seq samples 94,800 exons, 69,052 splice junctions, 73,325 transcripts and 41,204 genes in intergenic and antisense regions (Table 1b). These novel elements increase the Gencode collection of exons, splice sites, transcripts and genes by 19%, 22%, 45% and 80%, respectively. The increase in the number of genes and the relatively low contribution of novel splice sites is primarily caused by the detection of both polyadenylated and non-polyadenylated mono-exonic

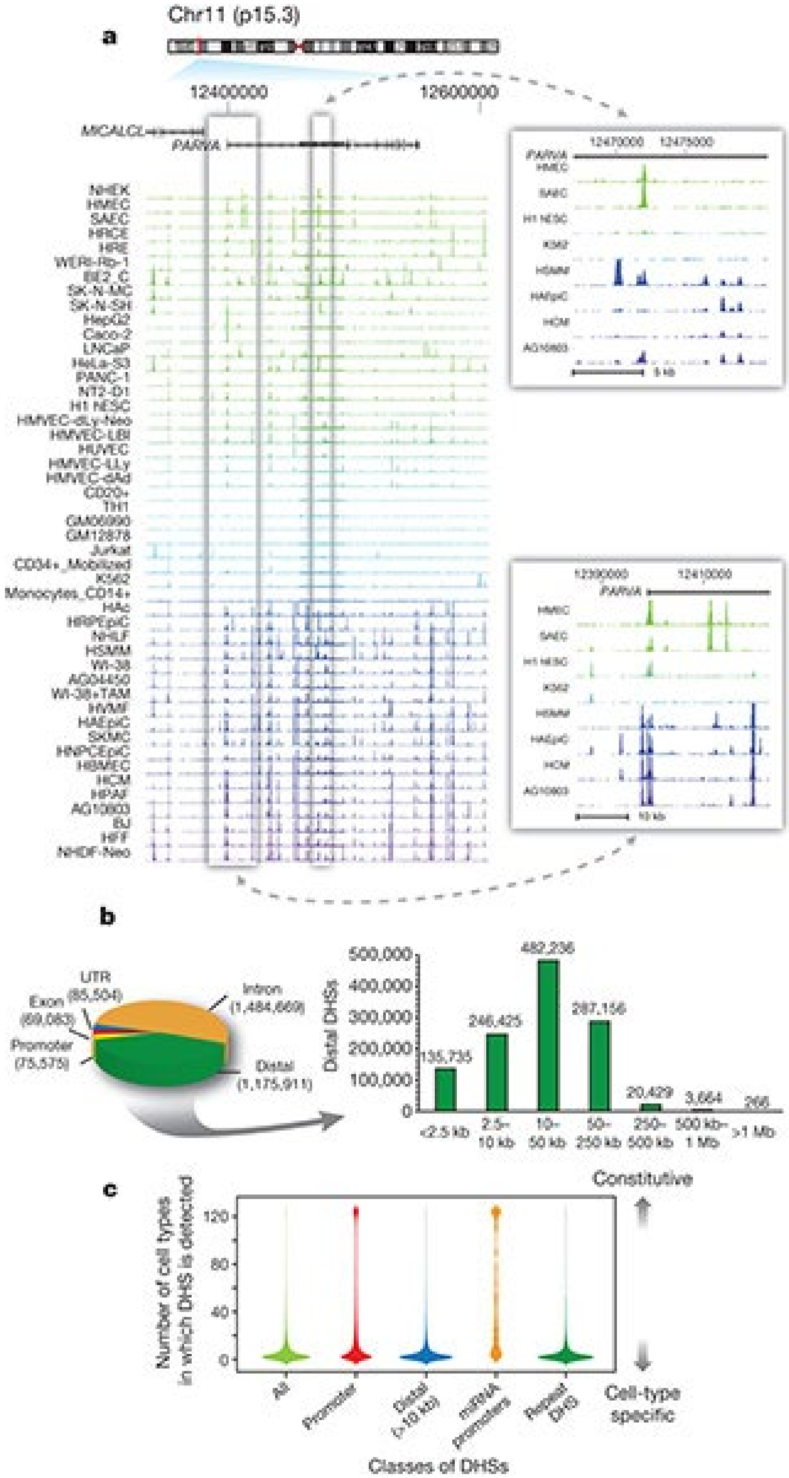


**Figure 1 | General features of the DHS landscape. (a) Density of DNase I cleavage sites for selected cell types, shown for an example ~350-kb region. Two regions are shown to the right in greater detail. (b) Left: distribution of 2,890,742 DHSs with respect to GENCODE gene annotations. Promoter DHSs are defined as the first DHS localizing within 1 kb upstream of a GENCODE TSS. Right: distribution of intergenic DHSs relative to Gencode TSSs. (c) Distributions of the number of cell types, from 1 to 125 (*y* axis), in which DHSs in each of four classes (*x* axis) are observed. Width of each shape at a given *y* value shows the relative frequency of DHSs present in that number of cell types.**

transcripts (Supplementary Table 3). Detection of unspliced transcripts could partially be an artefact caused by low levels of DNA contamination or by incomplete determination of transcript structures.

Independent validation of multi-exonic transcript models and the associated predicted coding products were carried out using overlapping targeted 454 Life Sciences (Roche) paired-end reads and mass spectrometry.

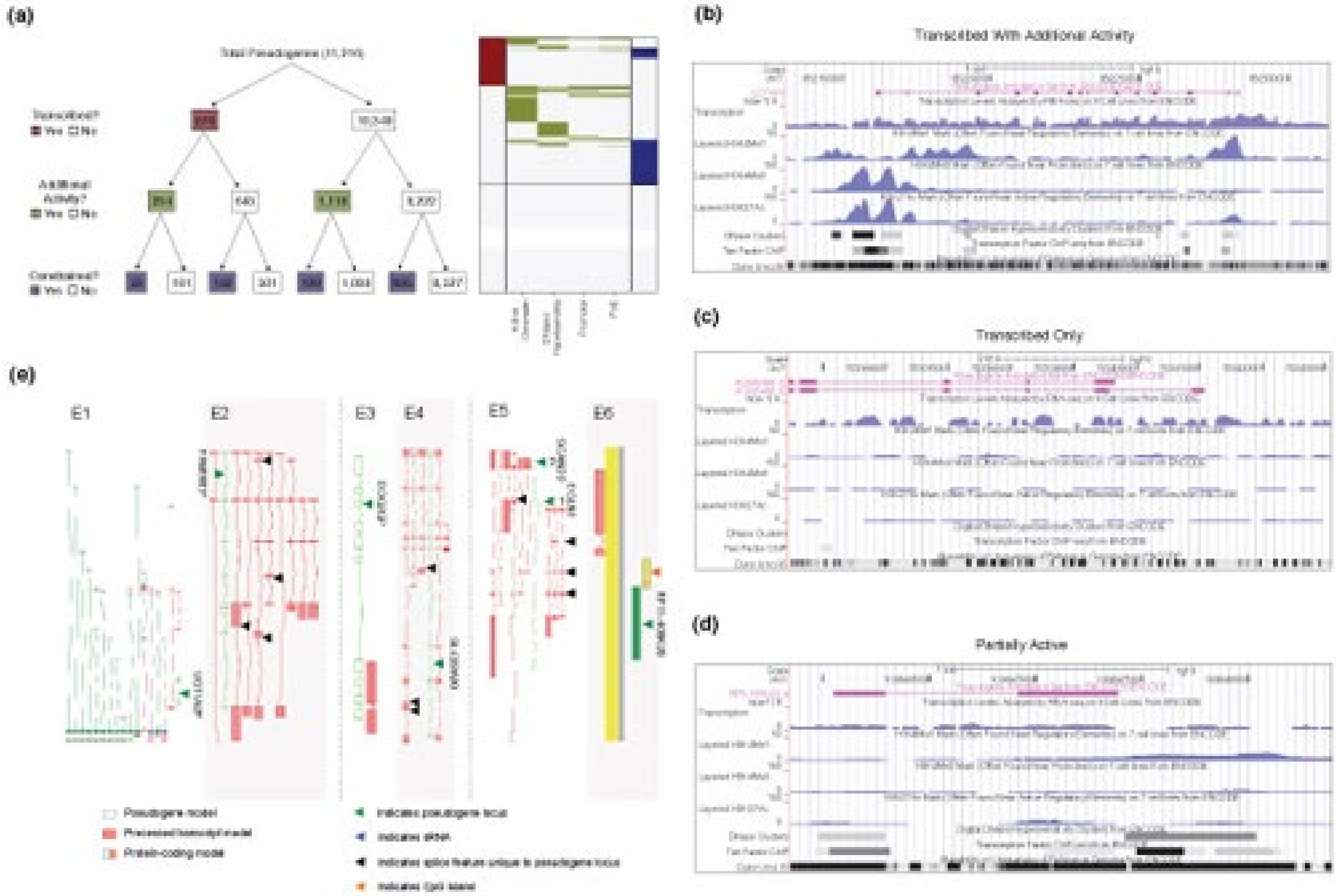


**Figure 12. | Summary of pseudogene annotation and case studies. (a) A heatmap showing the annotation for transcribed pseudogenes including active chromatin segmentation, DNaseI hypersensitivity, active promoter, active Pol2, and conserved sequences. Raw data was from the K562 cell line. (b) A transcribed duplicated pseudogene (ID: ENST00000434500.1, genomics location: chr7: 65216129-65228323) showing consistent active chromatin accessibility, histone marks, and TFBS in its upstream sequences. (c) A transcribed processed pseudogene (ID: ENST00000355920.3, genomic location: chr7: 72333321-72339656) with no active chromatin features or conserved sequences. (d) A non-transcribed duplicated psudogene showing partial activity patterns (ID: ENST00000429752.2, genomic location chr1: 109646053-109647388). (e) Examples of partially active pseudogenes. (e1) and (e2) are examples of duplicated pseudogenes. (e1) shows UGT1A2P (ENST00000454886), indicated by the green arrowhead. UTG1A2P is a non-transcribed pseudogene with active chromatin and it is under negative selection. Coding exons of protein-coding paralogous loci are represented by dark green boxes and UTR exons are filled red boxes. (e2) shows FAM86EP (ENST00000510506) as open green boxes, which is a transcribed pseudogene with active chromatin and upstream transcription factor and Pol2 binding sites. The transcript models associated with the locus are displayed as filled red boxes. Black arrowheads indicate features novel to the pseudogene locus. (e3) and (e4) show two unitary pseudogenes. (e3) shows DOC2GP (ID: ENST00000514950) as open green boxes, and transcript models associated with the locus are shown as filled red boxes. (e4) shows SLC22A20 (ID: ENST00000530038), again the pseudogene model is represented as open green boxes, transcript models associated with the locus are filled red boxes, and black arrowheads indicate features novel to the pseudogene locus. (e5) and (e6) show two processed pseudogenes. (e5) shows a pseudogene EGLN1 (ID: ENST00000531623) inserted into duplicated pseudogene SCAND2 (ID: ENST00000541103), which is a transcribed pseudogene showing active chromatin but no upstream regulatory regions as seen in the parent gene. The pseudogene models are represented as open green boxes, transcript models associated with the locus are displayed as filled red boxes, and black arrowheads indicate features novel to the pseudogene locus. (e6) shows a processed pseudogene RP11-409K20 (ID: ENST00000417984), as a filled green box, which has been inserted into a CpG island, indicated by an orange arrowhead.**

Of approximately 3,000 intergenic and antisense transcript models tested, validation rates from 70% to 90% were observed, depending on the number of reads and IDR score. In addition, these experiments led to the identification of more than 22,000 novel splice sites not previously detected, meaning an almost eightfold increase in detection compared to the sites originally detected with RNA-seq (Supplementary Fig. 6). Using mass spectrometric analyses, we investigated what fraction of the novel Cufflinks transcript models show evidence consistent with protein expression. We produced 998,570 spectra from two cell lines (K562 and GM12878; for details see ref. 16), and mapped them to a three-frame translation of the novel Cufflinks models (Supplementary Material). At a 1% false discovery rate (FDR), we identified 419 novel models with 5 or more spectral and/or 2 or more peptide hits, of which only 56 were intergenic or antisense to Gencode genes (Supplementary Table 4 and Supplementary Fig. 7). Thus, most novel transcripts seem to lack protein-coding capacity.

### The transcriptome of nuclear subcompartments

For the K562 cell line, we also analysed RNA isolated from three subnuclear compartments (chromatin, nucleolus and nucleoplasm; Supplementary 5). Almost half (18,330) of the Gencode (v7) annotated genes detected for all 15 cell lines (35,494) were identified in the analysis of just these three nuclear subcompartments. In addition, there were as many novel unannotated genes found in K562 subcompartments as there were in all other data sets combined (Supplementary Table 5 and Table 1b). For all annotated (Supplementary Table 5.1) or novel (Supplementary Table 5.2) elements, only a small fraction in each subcompartment was unique to that compartment (Supplementary Table 6).

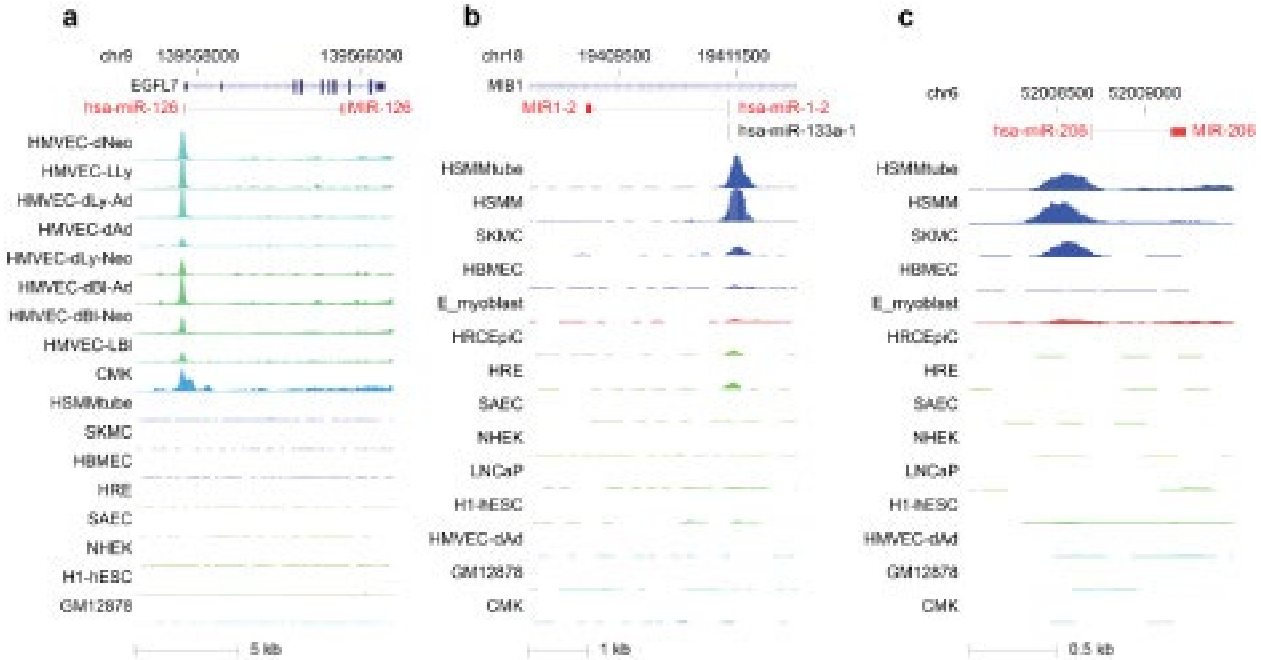
### Gene expression across cell lines

The analyses of RNAs isolated from different subcellular compartments also provide information concerning compartment-specific relative steady-state abundance and the post transcriptional processing state (spliced/unspliced, polyadenylated/non-polyadenylated, 5' capped/uncapped) for each of the detected transcripts. The observed range of gene expression spans six orders of magnitude for polyadenylated RNAs (from $10^{-2}$ to $10^{4}$ reads per kilobase per million reads (r.p.k.m.)), and five orders of magnitude (from $10^{-2}$ to $10^{3}$ r.p.k.m.) for non-polyadenylated RNAs (Fig. 3 and Supplementary Fig. 8a). The distribution of gene expression is very similar across cell lines, with protein-coding genes, as a class, having on average higher expression levels than long non-coding RNAs (lncRNAs). Assuming that 1-4 r.p.k.m. approximates to 1 copy per cell[19], we find that almost one-quarter of expressed protein-coding genes and 80% of the detected lncRNAs are present in our samples in 1 or fewer copies per cell. The general lower level of gene expression measured in lncRNAs may not necessarily be the result of consistent low RNA copy number in all cells within the population interrogated, but may also result from restricted expression in only a subpopulation of cells. In some cell lines, individual lncRNAs can exhibit steady-state expression levels as high as those of protein-coding genes. This is, for example, seen in the expression of the protein-coding gene actin, gamma 1 (*ACTG1*), and the non-coding gene, *H19* (Fig. 3). *ACTG1* transcripts are part of all non-muscle cytoskeleton systems within cells and show a steady-state expression level at the population level that is at least 1-2 logs greater than *H19*, a cytosolic non-coding RNA (ncRNA). However, when measured at the individual transcript level, expression of lncRNA transcripts is comparable to that of individual protein-coding transcripts (Supplementary Fig. 8b).

Novel antisense and intergenic genes predicted in this study comprise a third clustering of RNAs with levels of expression ranging from $10^{-4}$ to $10^{-1}$ r.p.k.m. As a class, only protein-coding genes seem to be enriched in the cytosol, making the nucleus a centre for the accumulation of ncRNAs (Fig. 3). Other gene classes, such as pseudogenes and small annotated ncRNAs, also show subcellular compartmental enrichment (Supplementary Fig. 9).

### Annotated small RNAs

Currently, a total of 7,053 small RNAs are annotated by Gencode, 85% of which correspond to four major classes: small nuclear (sn)RNAs, small nucleolar (sno)RNAs, micro (mi)RNAs and transfer (t)RNAs (Table 2a). Overall we find 28% of all annotated small RNAs to be expressed in at least one cell line (Table 2a).



**Supplementary Figure 3 | Accessible chromatin peaks overlapping microRNA promoters. Three examples of DHSs overlapping microRNA promoters. Peaks are usually observed in cell types consistent with known function of the microRNA. Panel (a) shows DNaseI signal at the promoter for MIR126. MIR126 is intronic, part of the transcript of the EGFL7 gene. MIR126 has a DHS at the promoter in several endothelial cell lines, consistent with its known function[1]. Panel (b) shows chromatin accessibility at the promoter for MIR1-2. The transcript is antisense of the MB1 gene. DHSs can be seen in muscle cell lines. Panel (c) shows a DHS at a potential promoter site in the muscle cell types HSMM, HSMMtube, SKMC, and myoblast. MIR1-2 and MIR206 are known to be involved in muscle function[2].**

The distribution of annotated small RNAs differs markedly between cytosolic and nuclear compartments (Supplementary Fig. 14a). We found that the small RNA classes were enriched in those compartments where they are known to perform their functions: miRNAs and tRNAs in the cytosol, and snoRNAs in the nucleus. Interestingly, snRNAs were equally abundant in both the nucleus and the cytosol. When specifically interrogating the subnuclear compartments of the K562 cell line, however, snRNAs seem to be present in very high abundance in the chromatin-associated RNA fraction (Supplementary Fig. 14b, c). This striking enrichment is consistent with splicing being predominantly co-transcriptional[17,26].

## Unannotated short RNAs

We detected two types of unannotated short RNAs. The first type corresponds to subfragments of annotated small RNAs. Because we performed 36-nucleotide end-sequencing of the small RNA fraction, we expected RNA-seq reads to map to the 5' end of the small RNAs. Supplementary Figure 15 shows the mapping profile of reads along small RNA genes. In both the nuclear and cytosolic compartments, we indeed detected accumulation of reads at the start of snoRNAs and at the guide and passenger sequences of annotated miRNAs. For snRNAs, however, we observed three prominent peaks: the expected one at the 5' end and two smaller ones at the middle and at the 3' end of the gene, indicating fragmentation of some snRNAs. Finally, tRNAs seem not to have any prominent sets of 5' end fragments present at levels greater than what is seen at the annotated 5' termini. Whereas subfragments of mature tRNAs have been reported previously, these reports were confined to distinct alleles of only a few tRNA genes[27-29].

The second and largest source of unannotated short RNAs corresponds to novel short RNAs (Table 2b) that map outside of annotated ones. Almost 90% of these are only observed in one cell line and are present at low copy numbers. Nearly 40% of these unannotated short RNAs are associated with promoter and terminator regions of annotated genes (promoter-associated short RNAs (PASRs) and termini-associated short RNAs (TASRs)), and their position relative to TSSs and transcription termination sites is similar to previous results[4].

### Analysing long non-coding transcript annotation

Over the last decade evidence from numerous high-throughput array experiments have indicated that evolution of the developmental processes regulating complex organisms can be attributed to the non-coding region and not only to the protein-coding region of the genome (Bertone *et al.* 2004; Mattick 2004; Kapranov *et al.* 2007; Clark *et al.* 2011). The GENCODE gene set has always attempted to catalogue this non-coding transcription utilizing a combination of computational analysis, human and mammalian cDNAs/ESTs alignments, and extensive manual curation to validate their non-coding potential. GENCODE 7 contains 9,640 long non-coding RNAs (lncRNAs) loci, representing 15,512 trancripts, which is the largest manually curated catalogue of human lncRNAs currently publicly available. All the lncRNA loci in the catalogue originate from the manual annotation pipeline and are initially classified as non-coding due to the lack of homology to any protein, no reasonable sized open reading frame (not subject to NMD) and no high conservation, confirmed by PhyloCSF (see later section), through the majority of exons. The transcripts are not required to be polyadenylated but 16.8% are, and chromatin marks have been identified for 13.9% (Derrien *et al.* 2012). These lncRNAs can been further reclassified into the following locus biotypes based on their location with respect to protein-coding genes:

1. **Antisense RNAs:** Locus that has at least one transcripts that intersect any exon of a protein-coding locus on the opposite strand, or published evidence of antisense regulation of a coding gene.
2. **LincRNA:** Locus is intergenic non-coding RNA loci.
3. **Sense overlapping:** Locus contains a coding gene within an intron on the same strand.
4. **Sense intronic:** Locus resides within intron of a coding gene, but does not intersect any exons on the same strand.
5. **Processed transcript:** Locus where non of its transcripts contain an open reading frame (ORF) and cannot be placed in any of the other categories because of complexity in their structure.

In summary the lncRNAs data set in GENCODE 7 consists of 5,058 lincRNA loci, 3,214 antisense loci, 378 sense intronic loci and 930 processed transcripts loci. Manually evaluating the RNA-seq models generated from HBM data and ENCODE data could potentially double this number in later releases of GENCODE and produce a uniform data set.

Here, we present and analyze the most comprehensive human lncRNA annotation to date, produced by the GENCODE consortium within the framework of the ENCODE project and comprising 9277 manually annotated genes producing 14,880 transcripts. Our analyses indicate that lncRNAs are generated through pathways similar to that of protein-coding genes, with similar histone-modification profiles, splicing signals, and exon/intron lengths. In contrast to protein-coding genes, however, lncRNAs display a striking bias toward two-exon transcripts, they are predominantly localized in the chromatin and nucleus, and a fraction appear to be preferentially processed into small RNAs. They are under stronger selective pressure than neutrally evolving sequences-particularly in their promoter regions, which display levels of selection comparable to protein-coding genes. Importantly, about one-third seem to have arisen within the primate lineage. Comprehensive analysis of their expression in multiple human organs and brain regions shows that lncRNAs are generally less expressed than protein-coding genes, and display more tissue-specific expression patterns, with a large fraction of tissue-specific lncRNAs expressed in the brain.

Expression correlation analysis indicates that lncRNAs show particularly striking positive correlation with the expression of antisense coding genes.

To estimate the fraction of lncRNAs that are translated *in vivo*, we compare the rate of detection of lncRNA translation to that for mRNAs expressed at similar levels. This is necessary, because otherwise any conclusions about the translational competency of lncRNAs would be subject to statistical confounding with levels and patterns of transcription. By interrogating our predictive models, we can "regress out" transcriptional effects on the detectability of peptides (see Methods for details). For mRNAs with expression levels comparable to those of lncRNAs, in GM12878, between 4.4% and 5.9% code for detected peptides (see Table 1 and Fig. 1). These numbers are directly comparable to the 0.33% of lncRNAs with detected translation in the same cell line (*p* < 10-16, two-sided Chi-square test). For K562 we have detected translation for between 1.5% and 1.8% of mRNAs with lncRNA-consistent expression patterns, and 0.09% of lncRNAs ($p < 10$-16, two-sided Chi-square test). Hence, lncRNAs are likely between 13- and 20-fold depleted for detected translation given their expression patterns. We can obtain an upper bound for the fraction of GENCODE v7 lncRNAs translated in vivo by considering that we "should have detected" peptides corresponding to 100% of mRNAs. This is an upper bound because clearly not all mRNAs are expressed in these cells, and hence cannot produce peptides. Indeed, we have zero expression values across all compartments for 5.5% of GENCODE v7 mRNAs in GM12878 and 6.0% in K562, 7.1% are zero across all polyA+ samples in GM12878, and 9.2% in K562, and finally 60% are zero in at least one compartment in GM12878 and 51% in K562. Under the conservative model that all mRNAs were detectable, we infer that at least 92% of GENCODE v7 lncRNAs are untranslated in these cell lines.

We have demonstrated that lncRNAs are depleted for peptides that are detectable in our tandem MS/MS assay, but it remains possible that extremely short or rapidly degraded polypeptides exist that have gone undetected. The length of ORFs is largely uncorrelated with the number of peptides we detected ($\rho \sim 0.08$, $r \sim 0.005$). In Supp. Fig. 5A we see that ORFs with detected peptides are enriched for long ORFs compared to the GENCODE v7 total ORF set (KS-2-sample test $p < 10$-16), but that this effect is dominated by an enrichment for ORFs of more than 3K amino acids, rather than by a depletion of short ORFs (Supp. Fig. 5B). However, the shortest GENCODE v7 ORF for which we have identified a peptide is 69 nucleotides, 23 amino acids; this implies an empirical size limit on detectability for our current data. Hence, it is possible that a population of short polypeptides has escaped the detection limits of our current MS/MS assay. Although we cannot rule out this possibility, we can provide an empirical bound: if the translation of short ORFs into stable polypeptides is wide-spread in the GENCODE v7 lncRNAs, then these likely encode polypeptides shorter than ~23 amino acids in length.

High-throughput immunogenomic analysis of human peptides indicates that they can function as novel autoantigens (Larman *et al.* 2011). Ectopic lncRNAs may represent one source of ORFs giving rise to such autoantigenic peptides. Accordingly, disease specificity and immunogenicity of lncRNA-encoded peptides may warrant future investigation, and may enhance ENCODE's impact on clinical and translational medicine.

### Partial Activity of Pseudogenes

We have integrated a large amount of genome-wide functional genomics data, together with expression and conservation data, to create a pseudogene annotation resource psiDR. This allows us to comprehensively examine pseudogenes activity from different perspectives such as transcription, regulation and evolution. We found a number of pseudogenes showing activity and more interestingly, a group of pseudogenes exhibiting various ranges of partial activity. Partially active pseudogenes were defined by a series of simple models based on transcription evidence, chromatin state, DNaseI hypersensitivity, upstream regulatory elements, and selection pressure. Different combinations of those features led to the characterization of pseudogenes as being partially active. One can speculate that partial activity may correspond to the process of resurrection of a pseudogene as a ncRNA or that it is in the process of dying and losing function. We believe that the various partially active pseudogenes provide a rich informative resource to aid understanding of pseudogene function and evolution.

One of the key aspects in defining the partially active pseudogenes is their upstream regulatory region. The presence or absence of regulatory elements is essential to understanding the evolutionary stage of the partially active pseudogenes. For example, a pseudogene showing active promoters and TFBS but lacking transcription evidence is believed to be a "dying" gene while a pseudogene with markedly different upstream elements compared to its parent gene but showing evidence of transcription is regarded as being potentially "resurrected". In the present paper we define the partially active pseudogenes based on a number of genomic features, namely transcription factor binding sites, histone marks, DNA accessibility, etc. However, we expect that future functional genomics datasets will complete the activity profile of pseudogenes. In particular integration of DNA methylation, nucleosome positioning, ChIA-PET, and HITS-CLIP datasets will provide a useful addition to the ENCODE pseudogene resource.

In conclusion, by integrating GENCODE pseudogene annotation, extensive functional genomics data from ENCODE and the variation data from the 1000 Genome project, we provide a comprehensive resource for pseudogene annotation and activity in the human genome. This resource has allowed us to classify pseudogenes with various attributes, which will enable interested researchers to identify expressed pseudogenes with potential function. Recent studies have shown the various ways by which pseudogenes regulate the expression of protein-coding genes and underscored the importance of identifying functional pseudogenes. We believe this resource provides data that can be used to further research in this direction. In particular, it is useful for understanding the regulatory role of pseudogenes, especially in cancer and other developmental processes. The comprehensive annotation of human pseudogenes also allows its comparison with pseudogenes from other model organisms such as mouse, worm, fly, and cress, which can provide valuable information on genome evolution.

### RT-PCR-seq to validate genome annotations

Secondly, we designed primer and selected for experimental validation by RT-PCR-seq a set of 10,162 different splice sites of 6,831 "novel" or "putative" GENCODE genes representing 9,213 distinct transcripts, as well as 486 ENCODE predicted models (e.g. in (Gotea *et al.* 2012)). Targeted splice sites can be divided into two classes: i) exon-exon junctions where one primer could be placed within 75 nucleotides of the junction ("Multi-span", Figure 1B)) and which will results in about half of the sequencing reads necessarily covering the junction (Figure 2A); ii) junctions where this was unfeasible ("Multi", Figure 1B) and in which sequencing reads will generally not reach the splice site. We identified, however a high number of sequencing reads crossing the targeted splice even for this category of amplimers (Figure 2A). A third set of models is formed by the monoexonic transcripts ("Mono", all belonging to the set of ENCODE predicted models; Figure 1B). As models belonging to this category are sensitive to genomic DNA contaminations, they were assessed by amplification of cDNA in which a dNTP analog was incorporated as described in (Washietl *et al.* 2007) (see Methods). Models belonging to each of these three categories were considered experimentally validated with different criteria as described in Methods and summarized in Figure 1C. We performed RT-PCR-seq for these 10,648 target loci and produced a total of 1,845,687,068 reads (Supplementary Table S1). 78% of them passed the quality threshold (mean Phred quality score $\geq$ 23) and were mappable on the genome and/or the GENCODE transcriptome using Bowtie (Langmead *et al.* 2009). The overall validation rate across all tissues is extremely high, reaching between 73 and 87% for each biotype/status combinations tested (Figure 2B, Supplementary Table S2-S3). Examples of validated "Multi-span" and monoexonic transcript models are presented in Supplementary Figure S1A-B. The transcriptome of testis showed the highest complexity (highest percentage of validated transcript models at 55%, Figure 2D) in accordance with previous reports and consolidating the view that chromatin is more relaxed in this tissue, leading to higher transcriptional activity (Denoeud *et al.* 2007). Each biotype/status combination is validated at comparable rates in the different investigated tissues with the exception of putative processed transcripts. Processed transcripts (putative and novel) are mainly identified in testis and are significantly more tissue-specific than other biotypes ($P = 1.52 \times 10^{-83}$, Fisher exact test; Figure 2C). This is consistent with the hypothesis that a large fraction of GENCODE putative processed transcript loci correspond to long non-coding RNAs genes (lncRNAs), which are expressed at lower levels and more specifically than other genes (Derrien *et al.* 2012). Our ability to validate a large fraction of less well-supported GENCODE gene models further emphasizes the extremely high quality reached by the GENCODE gene set originating from the manual annotation involved.

### RT-PCR-seq to substantiate RNA-seq predictions

Since we showed that GENCODE (or any other annotation for that matter) does not yet fully represent the complexity of the human transcriptome, we took advantage of the deep transcriptome profiling by HBM to uncover novel gene models. The 3.8 billion individual sequence reads were aligned on the human genome to predict alignment blocks (rough exon models), splice sites and finally novel gene models (see Methods for the Ensembl RNA-seq pipeline). At each locus, the transcript model with the greatest number of supporting reads is displayed on the Ensembl genome browser. 5918 of them do not overlap any loci depicted in GENCODE freeze version 7. Thus they potentially represent new non-coding RNA genes or alternatively unannotated 5′ or 3′UTR portions of known genes, as the vast majority of these models were shown to have poor coding potential using comparative genomics and mass spectrometry (Lin *et al.* 2011; Harrow *et al.* 2012). We could design primers on splice-junctions of 1601 of those models to assess them experimentally by RT-PCR-seq. We validated 73% of the new HBM models outlined by the Ensembl predictions in an average of 4.5 tissues (Figure 2B and 2C), *de facto* enriching the future complexity of the GENCODE annotation of non-coding RNAs genes by 1168 novel genes, a 3.7% increase. As this rate of validation is close to the sensitivity of the RT-PCR-seq method with 8 tissues for non-coding transcripts (79%, see above) we suggest that a large fraction of the non-validated HBM models might be *bona fide* transcripts rather than false positive predictions. Our findings demonstrate the effectiveness of RNA-seq combined with RT-PCR-seq to uncover new genome features. These two technologies were simultaneously similarly paired to unravel expressed pseudogenes by the GENCODE consortium (Pei *et al.* 2012).

Large ongoing efforts to profile the human transcriptome by RNA-seq, (e.g. HBM and ENCODE) confirm this assumption as they revealed a large number of transcribed islands that do not overlap GENCODE annotations (Djebali *et al.* 2012b). These isolated islands and archipelagos potentially represent novel exons and novel genes, respectively. Our RT-PCRseq method has proven to be a powerful tool to evaluate such *de novo* transcript models derived from RNA-seq experiments. We validated 72% of the assessed model confirming that they are *bona fide* transcribed units missing from the current annotations. Together these observations support the notion that the human genome is pervasively transcribed as suggested by multiple authors (Kapranov *et al.* 2002; The ENCODE Project Consortium 2007; Denoeud *et al.* 2007; Djebali *et al.* 2008; Clark *et al.* 2011; Djebali *et al.* 2012b).

MicroRNAs (miRNAs) comprise a major class of regulatory molecules and have been extensively studied, resulting in consensus annotation of hundreds of conserved miRNA genes[11], approximately one-third of which are organized in polycistronic clusters[12]. However, most predicted promoters driving microRNA expression lack experimental evidence. Of 329 unique annotated miRNA TSSs (Supplementary Methods), 300 (91%) either coincided with or closely approximated (<500 base pairs (bp)) a DHS. Chromatin accessibility at miRNA promoters was highly promiscuous compared with Gencode TSSs (Fig. 1c, fourth column), and showed cell lineage organization, paralleling the known regulatory roles of well-annotated lineage-specific miRNAs (Supplementary Fig. 3).