HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2013 June 17.

Published in final edited form as:

Nature. 2012 September 6; 489(7414): 101–108. doi:10.1038/nature11233.

Landscape of transcription in human cells

A full list of authors and affiliations appears at the end of the article.

Summary

Eukaryotic cells make many types of primary and processed RNAs that are found either in specific sub-cellular compartments or throughout the cells. A complete catalogue of these RNAs is not yet available and their characteristic sub-cellular localizations are also poorly understood. Since RNA represents the direct output of the genetic information encoded by genomes and a significant proportion of a cell's regulatory capabilities are focused on its synthesis, processing, transport, modifications and translation, the generation of such a catalogue is crucial for understanding genome function. Here we report evidence that three quarters of the human genome is capable of being transcribed, as well as observations about the range and levels of expression, localization, processing fates, regulatory regions and modifications of almost all currently annotated and thousands of previously unannotated RNAs. These observations taken together prompt to a redefinition of the concept of a gene.

As the technologies for RNA profiling and for cell type isolation and culture continue to improve, the catalogue of RNA types has grown and led to an increased appreciation for the numerous biological roles played by RNA, arguably putting them on par with the functional importance of proteins[1]. The Encyclopedia of DNA Elements (ENCODE) project has sought to catalogue the repertoire of RNAs produced by human cells as part of the intended goal of identifying and characterizing the functional elements present in the human genome sequence[2]. The pilot phase of the ENCODE project[3] examined approximately 1% of the human genome and observed that the gene-rich and gene-poor regions were pervasively transcribed, confirming results of prior studies[4,5]. During the second phase of the ENCODE project, the scope of examination was broadened to interrogate the complete human genome.

Corresponding Authors: - Thomas R. Gingeras, Cold Spring Harbor Laboratory. gingeras@cshl.edu - Roderic Guigó, Centre for Genomic Regulation. roderic.guigo@crg.eu.

*These authors contributed equally to this work

Author Information A complete set of data files can be downloaded at GEO under the following accessions: GSE26284 (CSHL, Long RNA), GSE33480 (Caltech, A+ RNA-seq) GSE24565 (CSHL, Short RNA), GSE33600 (GIS, RNA-PET), GSE34448 (RIKEN, CAGE) or viewable at the UCSC Genome Browser at http://genome-preview.ucsc.edu/ENCODE/. Reprints and permissions information is available at www.nature.com/reprints.

Method summary: see Supplementary Material

Author Contributions

Lead the project and oversaw the analysis: T.R.G., R.G., P.C., B.W., Y.R., M.C.G., G.H., S.E.A., A.R., T.H., M.G., Y.H. Oversaw or significantly contributed to data generation: C.A.D., X.R., B.A.W., P.C., Major contributions towards data processing and analysis: S.D., A.M., A.D., T.L., A.M.M., A.T., J.L., W.L., F.S., C.X., G.K.M., J.K., C.Z., J.R., M.R., F.K., J.H. **Data production and analysis**: R.F.A., T.A., I.A., M.T.B., N.S.B., P.B., K.B., I.B., S.C., X.C., J.C., J.C., T.D., J.D., E.D., J.D., R.D., E.F., M.F., K.F-T., P.F., S.F., M.J.F., H.G., D.G., A.G., H.G., C.H., S.J., R.J., P.K., B.K., C.K., O.J.L., E.P., K.P., J.B.P., P.R., B.R., D.R., M.S., L.S., H-H. S., A.S., J.S., A.M.S., H.T., H.T., D.T., N.W., H.W., J.W., Y.Y. **Wrote the manuscript with input from authors**: T.R.G. and R.G.

Thus, we have sought to both provide a genome-wide catalogue of human transcripts and to identify the sub-cellular localization for the RNAs produced. Here we report identification and characterization of annotated and novel RNAs that are enriched in either of the two major cellular sub-compartments (nucleus and cytosol) for all 15 cell lines studied, and in three additional sub-nuclear compartments in one cell line. In addition, we have sought to determine if identified transcripts are modified at their 5′ and 3′ termini by the presence of a 7-methyl guanosine cap or polyadenylation, respectively. We further studied primary transcript and processed product relationships for a large proportion of the previously annotated long and small RNAs. These results considerably extend the current genome-wide annotated catalogue of long polyadenylated and small RNAs collected by the Gencode annotation group[6-8]. Taken together our genome-wide compilation of subcellular localized and product-precursor related RNAs serves as a public resource and reveals new and detailed facets of the RNA landscape:

- Cumulatively, we observed a total of 62.1% and 74.7% of the human genome to be covered by either processed or primary transcripts respectively, with no cell line showing more than 56.7% of the union of the expressed transcriptomes across all cell lines. The consequent reduction in the length of “intergenic regions” leads to a significant overlapping of neighboring gene regions and prompts a redefinition of a gene.
- Isoform expression by gene does not follow a minimalistic expression strategy resulting in a tendency for genes to express many isoforms simultaneously with a plateau at about 10-12 expressed isoforms per gene per cell line.
- Cell type-specific enhancers are promoters that are differentiable from other regulatory regions by the presence of novel RNA transcripts, chromatin marks and DNAse l hypersensitive sites.
- Coding and non-coding transcripts are predominantly localized in the cytosol and nucleus respectively, with a range of expression spanning six orders of magnitude for polyadenylated RNAs, and five orders of magnitude for non-polyadenylated RNAs.
- Approximately 6% of all annotated coding and non-coding transcripts overlap with small RNAs and are likely precursors to these small RNAs. The sub-cellular localization of both annotated and unannotated short RNAs is highly specific.

RNA dataset generation

We performed sub-cellular compartment fractionation (whole cell, nucleus and cytosol) prior to RNA isolation in 15 cell lines (Table S1) to deeply interrogate the human transcriptome. For the K562 cell line, we also performed additional nuclear sub-fractionation into: chromatin, nucleoplasm and nucleoli. The RNAs from each of these sub-compartments were prepared in replica and were separated based on length into >200 nucleotides (nt) (long) and <200 nt (short). Long RNAs were further fractionated into polyadenylated and non-polyadenylated transcripts. A number of complementary technologies were employed to characterize these RNA fractions as to their sequence (RNA-seq), sites of initiation of transcription (Cap-Analysis of Gene Expression -CAGE[9]) and

sites of 5′ and 3′ transcript termini (Paired End Tags -PET[10], Figure S1). Sequence reads were mapped and post-processed using a variety of software tools (Table S2, Figure S2). We used the mapped data to assemble and quantify *de novo* elements (exons, transcripts, genes, contigs, splice junctions and transcription start sites, TSS) as well as to quantify annotated Gencode (v7) elements. Elements and quantifications were further assessed for reproducibility between replicates using a non-parametric version (npIDR, Supplementary Material) of the Irreproducible Detection Rate (IDR) statistical test[11]. Only elements deemed to be reproducible with at least 90% likelihood were used in most analyses. The raw data, mapped data and elements were then made available by the ENCODE Data Coordination Center or DCC (http://genome.ucsc.edu/ENCODE/dataSummary.html) (Figure S2). These data, as well as additional data on all intermediate processing steps are available on the RNA Dashboard: http://genome.crg.cat/encode_RNA_dashboard/.

Long RNA expression landscape

Detection of annotated and novel transcripts

The Gencode gene (Figure S3a) and transcript (Figure S3b) 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 (Figure 1, and Table 1.1). 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 (Figure 1, width of box plots). Consistent with earlier studies, most annotated elements are present in both polyadenylated (Table S3a) and non-polyadenylated (Table S3b) samples[12-15]. 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 (Figure S4). Overall, the unique contribution of each cell line to the coverage of the genome tend to be small and similar for each cell line (Figure S5). We used the Cufflinks algorithm (see Supplementary Material), 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 1.2). 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-polyadenlyated mono-exonic transcripts (Table S3). Detection of unspliced transcripts could partially be an artifact, 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. 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 8-fold increase in detection compared to the sites originally detected with RNA-seq (Figure S6). 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 Khatun et al.[16]), and mapped them to a 3-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 (Table S4 and Figure S7). Thus, most novel transcripts appear to lack protein coding capacity.

The transcriptome of nuclear sub-compartments

For the K562 cell line, we also analyzed RNA isolated from three sub-nuclear compartments (chromatin, nucleolus and nucleoplasm, Table S5). 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 sub-compartments. In addition, there were as many novel unannotated genes found in K562 sub-compartments as there were in all other datasets combined (Table S5 vs. Table 1.2). For all annotated (Table S5.1) or novel (Table S5.2) elements, only a small fraction in each sub-compartment was unique to that compartment (Table S6).

The interrogation of different sub-cellular RNA fractions provides snapshots of the status of the RNA population along the RNA processing pathway. Thus, by analyzing short and long RNAs in the different sub-cellular compartments, we confirm that splicing predominantly occurs during transcription. By using RNA-seq to measure the degree of completion of splicing (Figure 2a), we observed that around most exons, introns are already being spliced in chromatin-associated RNA—the fraction that includes the RNAs in the process of being transcribed (Figure 2b). Concomitantly, we found strong enrichment specifically of spliceosomal small nuclear RNAs (snRNAs) in this RNA fraction (see short RNA expression landscape section below). Co-transcriptional splicing provides an explanation for the increasing evidence connecting chromatin structure to splicing regulation, and we have indeed observed that exons in the process of being spliced are enriched in a number of chromatin marks[17,18].

Gene expression across cell lines

The analyses of RNAs isolated from different sub-cellular 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 [RPKM]), and five orders of magnitude (from 10^{-2} to 10^{3} RPKM) for non-polyadenylated RNAs (Figure 3 and figure S8a). 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 RPKM 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* (Figure 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 ncRNA. However, when measured at the individual transcript level, expression of lncRNA transcripts is comparable to that of individual protein coding transcripts (Figure S8b).

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} RPKM. As a class, only protein coding genes appear enriched in the cytosol, making the nucleus a center for the accumulation of non-coding RNAs (Figure 3). Other gene classes, such as pseudogenes and small annotated ncRNAs, also show sub-cellular compartmental enrichment (Figure S9).

Higher variability and lower pairwise correlation of expression across all cell lines is consistent with lncRNAs contributing more to cell line specificity than protein-coding genes. Indeed, a considerable fraction (29%) of all expressed lncRNAs are detected in only one of the cell lines studied when considering the whole cell polyadenylated RNAs, while only 10% were expressed in all cell lines. Conversely, while a large fraction (53%) of expressed protein coding genes were constitutive (expressed in all cell lines), only ~7% were cell-line specific (Table S7, Figure S10).

Patterns of splicing

The analysis of the expression of alternative isoforms resulted in several observations. First, isoform expression does not seem to follow a minimalistic strategy. Genes tend to express many isoforms simultaneously, and as the number of annotated isoforms per gene grows, so does the number of expressed isoforms (Figure 4a). The increase, however, is not linear and appears to plateau at about 10-12 expressed isoforms per gene. We cannot obviously distinguish, however, whether this is the result of multiple isoforms expressed in the same cell or of different isoforms expressed in different cells within the interrogated population. Second, alternative isoforms within a gene are not expressed at similar levels, and one isoform dominates in a given condition—usually capturing a large fraction of the total gene expression (at least 30% even for genes with many isoforms, Figure 4b). Third, about three quarters of protein coding genes have at least two different dominant/major isoforms depending on the cell line (Figure S11a). Fourth, the number of major isoforms per gene grows with the number of annotated isoforms; indeed, the proportion of genes with *n* isoforms that express only one major isoform is strikingly proportional to *1/n* (Figure S11b). Fifth, variability of gene expression contributes more than variability of splicing ratios to the variability of transcript abundances across cell lines (Supplementary Material).

Alternative transcription initiation and termination

Based on RNA-seq analysis of polyadenylated RNAs, a total of 128,021 TSS were detected across all cell lines, of which 97,778 were previously annotated and 30,243 were novel intergenic/antisense TSS (Table S3a). CAGE tags, filtered by a hidden Markov model (HMM) based algorithm to differentiate between 5′ capped termini of polymerase II transcripts and recapping events[20] (Supplementary Material), identified a total of 82,783 non-redundant TSS (Table S8). Approximately 48% of the CAGE identified TSS are located within 500 bp of an annotated RNA-seq detected Gencode TSS, while an additional 3% are within 500 bp of a novel TSS (Figure S12). Interestingly, only ~72% of all CAGE sequencing reads map to TSS, indicating that the remaining 30% may originate from recapping events or from a new class of TSS.

Using data collected within the ENCODE consortium[21], we carried out a comparison of the Gencode/RNA-seq and CAGE determined TSSs and correlated them to chromatin and DNA features characteristic of initiation of transcription, such as DNAse hypersensitivity[22], chromatin modification and DNA binding elements[23,24]. All Gencode/RNA-seq determined TSS were examined in each of the cell lines (column 1, Figure S13). Of these redundant positions, 44.7% (199,146) of the RNA-seq supported TSS also displayed evidence of CAGE. Approximately half of these TSS positions are associated with at least one of the other characteristic features of transcription initiation (DNAse I, H3K27Ac and H3K4me3 chromatin modifications). Thus only a small minority of the TSS identified by either CAGE or RNA-seq/Gencode displayed all of the characteristics of the start of transcription (presence of DNAseI, H3K4me3, H3K27ac sites and either Taf1 or Tbp binding). This is consistent with the possibility that regulatory regions proximal to TSS, are of more than one type.

On the other hand, a total of 128,824 sites mapping within annotated Gencode transcripts were identified as potential sites of polyadenylation after trimming unmapped RNA-seq reads with long terminal polyadenine stretches[25]. About 20% of these mapped proximal to annotated polyadenylation sites (PAS) while the remaining 80% correspond to novel PAS of annotated genes, raising the average number of PAS per gene from 1.1 to 2.5. Generally, we observed a cell type preference for proximal PAS (closest to the annotated stop codon) in the cytosol compared to the nucleus (Supplementary Material).

Short RNA expression landscape

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). The distribution of annotated small RNAs differs markedly between cytosolic and nuclear compartments (Figure S14a). 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 sub-nuclear compartments of the K562 cell line, however, snRNAs appear to be present in very high abundance in the chromatin-associated RNA fraction (Figure S14bc). 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 sub-fragments of annotated small RNAs. Since we performed 36 nt end-sequencing of the small RNA fraction, we expected RNA-seq reads to map to the 5′ end of the small RNAs. Figure S15 shows the mapping profile of reads along small RNA genes. In both the nuclear and cytosolic compartments, we indeed detect 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, suggesting fragmentation of some snRNAs. Finally, tRNAs appear not to have any prominent sets of 5′ end fragments present at levels greater than what is seen at the annotated 5′ termini. While sub-fragments 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 correspond 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], termini associated short RNAs [TASRs]), and their position relative to TSS and transcription termination sites is similar to previously found[4].

Genealogy of short RNAs

Genome wide, 27% of annotated small RNAs reside within 8% of protein-coding and 5% within 3% of lncRNA genes (Figure S16). Overall, about 6% of all annotated long transcripts overlap with small RNAs and are likely precursors to these small RNAs. While the majority of these small RNAs reside in introns, when controlling for relative exon/intron length, we found that exons from lncRNAs are comparatively enriched as hosts for snoRNAs (Figure S17a). Additionally, 8.4% of Gencode annotated small RNAs map within novel intergenic transcripts with the majority overlapping annotated tRNAs. The enrichment for tRNAs was mostly in novel intergenic transcripts derived from non-polyadenylated RNAs (Figure S17b). Many long RNAs, both novel and annotated, thus appear to have dual roles, as functional (protein coding) RNAs, and as precursors for many important classes of small RNAs. Using RNA-seq data from K562, we investigated the preferential cellular localization of these RNA precursors (Figure S18). For mature miRNAs and tRNAs (cytosolic enrichment), the potential RNA precursors, identified as RNA-seq contigs overlapping the small RNAs, were detected to be predominantly nuclear (FigureS18a,d). Interestingly, while mature snRNAs were both nuclear and cytosolic, the overlapping long RNAs were observed to be primarily nuclear (Figure S18c). Finally, for snoRNAs (nuclear enrichment), potential long RNA precursors were decidedly observed to be both nuclear and

cytosolic (Figure S18b). Unannotated short RNAs were found overall not to be enriched in either the nuclear or cytosolic compartment (Figure S18e).

RNA editing and allele-specific expression

The sequence of transcripts can differ from the underlying genomic sequence as the result of post-transcriptional editing. We developed a pipeline to filter sequencing artifacts and identify genes that are RNA edited[30]. Focusing first on GM12878, a cell line that has been deeply resequenced, we find a total 51,557 RNA consistent single nucleotide variants within genic boundaries, 65% of which are present in dbSNP. Of the remainder, 1,186 SNVs in 430 genes (Figure S19a) survive our most stringent filters and 88% of these are candidate adenosine to inosine A->G(I) changes. Notably the next highest frequency of SNVs are for T->C (5%) and are primarily in regions with detectable antisense transcription[30]. We find similar A->G(I) frequencies of 75-84%, in 7 additional cell lines (Figure S19b). The remaining non-canonical edits amount to very few events in each cell line and are relatively evenly distributed (G->A is the third highest). These results do not support a recent report of a substantial number of non-canonical SNV edits in the RNA of human lymphoblastoid cells[31].

Using the AlleleSeq pipeline[32] on the SNPs in the GM12878 genome, we found that approximately 18% of both Gencode annotated protein coding and long non-coding genes exhibit allele-specific expression (ASE). The proportion of genes with ASE was similar in the three investigated RNA fractions (whole-cell, cytoplasm and nucleus, Table S9 and Supplementary Material).

Repeat region transcription

About 18% (14,828) of CAGE defined TSS regions overlap repetitive elements. More precisely, we find 322, 315, 507 and 1,262 intergenic CAGE clusters overlapping LINE, SINE, LTR and other repeat elements respectively (see Supplementary Material). Measuring Shannon entropy across cell lines, we found that CAGE clusters mapping to repeat regions were noticeably more narrowly expressed that CAGE clusters mapping within genic regions (Figure S20a). We represented the correlation of levels of expression compared to cell types as heat maps drawn separately for each of the three repeat element families (LINE, SINE and LTR) (Figure S20b-d). While a large proportion of the transcripts in the human genome are thought to be initiated from repetitive elements (especially retrotransposon elements[33]), these data clearly point to cell line specificity as the main characteristic of transcripts emanating from repeat regions.

Characterization of enhancer RNA

It has recently been reported that RNA polymerase II binds some distal enhancer regions and can produce enhancer-associated transcripts named eRNA[34-36]. We used our RNA assays to detect and characterize transcriptional activity at enhancer loci predicted genome-wide from ENCODE ChIP-seq data [21,37].

Figure 5a shows the aggregate pattern of RNA-seq and CAGE signal in a strand specific manner around the subset of predicted gene-distal enhancers containing DNAse I hypersensitive sites and centered on those sites. In these plots, as denoted by the accumulation of CAGE tags signifying transcription start sites (TSS), transcription initiation within the enhancer region is observed, and continues outwards for several kilobases. This behaviour can be observed for the polyadenylated and non-polyadenylated RNA fractions mapping in both intronic and intergenic regions. As previously reported[34], we observe a large diversity of expression levels at each of the transcribed enhancers. Polyadenylated to non-polyadenyated RNA ratios, as well as nuclear to cytoplasmic ratios vary at individual enhancers (Figure S21ab). However, contrary to some previous reports, while the majority of eRNAs are prevalent in the nuclear non-polyadenylated RNA fraction, some eRNAs appeared to be polyadenylated in the nucleus. This pattern was significantly different compared to transcripts from Gencode annotated and novel predicted[21] promoters (Figure 5b).

Transcribed enhancers on average show a significantly different pattern of chromatin modifications than non transcribed ones[38-41]. The enhancer regions displayed stronger signals for H3K4 methylation, H3K27 acetylation and H3K79 dimethylation along with higher levels of RNA polymerase II binding, all associated with transcriptional initiation and elongation (Figure 5c). Both the transcripts and the chromatin states are cell-type specific (Figure 5d). Taking the GM12878 cell line as an example, the enhancer loci producing eRNA demonstrate enrichment of CAGE tag detection (Figure 5d.1) and the presence of H3K27ac histone modification (Figure 5d.2) in this cell line compared to five other analyzed cell lines. This strongly suggests that the regulatory regions governing the expression of enhancer transcripts are distinguished from regulatory regions located at the beginning of genic regions.

Conclusion: Genome-wide coverage of transcribed regions of the human genome and its consequences

The cumulative coverage of transcribed regions in the 15 cell lines across the human genome is 62.1% and 74.7% for processed and primary transcripts (Table S10 and Figure S22). On average for each cell line, 39% of the genome is covered by primary transcripts, and 22% by processed RNAs. No cell line showed transcription of more than 56.7% of the union of the expressed transcriptomes across all cell lines. When mapping the current RNA-seq data to the ENCODE pilot regions (Table S10), we observed a similar, albeit higher, extent of transcriptional coverage of 73.3% for processed RNAs, and 84.5% for primary transcripts. Previously reported estimates in these regions for processed and primary transcripts, were 24% and 93% respectively (Table S2.4.3[3]). The increased genome coverage by processed RNAs stems largely from the inclusion of non-polyadenylated RNAs in the current study. Other than that, given the differences in the samples studied, the selection of pilot regions with high genic content, the increase of annotated genomic regions over time, and the different technologies used to interrogate transcription, both estimates are in reasonable agreement.

As a consequence of both the expansion of genic regions by the discovery of new isoforms and the identification of novel intergenic transcripts, there has been a marked increase in the number of intergenic regions (from 32,481 to 60,250) due to their fragmentation and a decrease in their lengths (from 14,170bp to 3,949bp median length, Figure 6). Concordantly, we observe an increased overlap of genic regions. Since the determination of genic regions is currently defined by the cumulative lengths of the isoforms and their genetic association to phenotypic characteristics, the likely continued reduction in the lengths of intergenic regions will steadily lead to the overlap of most genes previously assumed to be distinct genetic loci. This supports and is consistent with earlier observations of a highly interleaved transcribed genome[12], but more importantly, prompts the reconsideration of the definition of a gene. Being this a consistent characteristic of annotated genomes, we would propose that the transcript be considered as the basic atomic unit of inheritance. Concomitantly, the term gene would then denote a higher order concept intended to capture all those transcripts (eventually divorced from their genomic locations) that contribute to a given phenotypic trait.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Sarah Djebali[1,*], Carrie A. Davis[2,*], Angelika Merkel[1], Alex Dobin[2], Timo Lassmann[7], Ali M. Mortazavi[5,8], Andrea Tanzer[1], Julien Lagarde[1], Wei Lin[2], Felix Schlesinger[2], Chenghai Xue[2], Georgi K. Marinov[5], Jainab Khatun[4], Brian A. Williams[5], Chris Zaleski[2], Joel Rozowsky[13,14], Maik Röder[1], Felix Kokocinski[12], Rehab F. Abdelhamid[7], Tyler Alioto[1], Igor Antoshechkin[5], Michael T. Baer[2], Nadav S. Bar[17], Philippe Batut[2], Kimberly Bell[2], Ian Bell[3], Sudipto Chakrabortty[2], Xian Chen[11], Jacqueline Chrast[10], Joao Curado[1], Thomas Derrien[1], Jorg Drenkow[2], Erica Dumais[3], Jacqueline Dumais[3], Radha Duttagupta[3], Emilie Falconnet[9], Meagan Fastuca[2], Kata Fejes-Toth[2], Pedro Ferreira[1], Sylvain Foissac[3], Melissa J. Fullwood[6], Hui Gao[3], David Gonzalez[1], Assaf Gordon[2], Harsha Gunawardena[11], Cedric Howald[10], Sonali Jha[2], Rory Johnson[1], Philipp Kapranov[3,16], Brandon King[5], Colin Kingswood[1], Oscar J. Luo[6], Eddie Park[8], Kimberly Persaud[2], Jonathan B. Preall[2], Paolo Ribeca[1], Brian Risk[4], Daniel Robyr[9], Michael Sammeth[1], Lorian Schaffer[5], Lei-Hoon See[2], Atif Shahab[6], Jorgen Skancke[1,17], Ana Maria Suzuki[7], Hazuki Takahashi[7], Hagen Tilgner[1], Diane Trout[5], Nathalie Walters[10], Huaien Wang[2], John Wrobel[4], Yanbao Yu[11], Xiaoan Ruan[6], Yoshihide Hayashizaki[7], Jennifer Harrow[12], Mark Gerstein[13,14,15], Tim Hubbard[12], Alexandre Reymond[10], Stylianos E. Antonarakis[9], Gregory Hannon[2], Morgan C. Giddings[4,11], Yijun Ruan[6], Barbara Wold[5], Piero Carninci[7], Roderic Guigó[1], and Thomas R. Gingeras[2,3]

Affiliations

[1]Centre for Genomic Regulation (CRG) and UPF, Doctor Aiguader, 88 . Barcelona, Catalunya, Spain 08003.

[2]Cold Spring Harbor Laboratory, Functional Genomics, 1 Bungtown Rd. Cold Spring Harbor, NY, USA 11742.

[3]Affymetrix, Inc, 3380 Central Expressway, Santa Clara, CA. USA 95051.

[4]Boise State University, College of Arts & Sciences, 1910 University Dr. Boise, ID USA 83725.

[5]California Institute of Technology, Division of Biology, 91125. 2 Beckman Institute, Pasadena, CA USA 91125.

[6]Genome Institute of Singapore, Genome Technology and Biology, 60 Biopolis Street, #02-01, Genome, Singapore, Singapore 138672.

[7]RIKEN Yokohama Institute, RIKEN Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa Japan 230-0045.

[8]University of California Irvine, Dept of. Developmental and Cell Biology, 2300 Biological Sciences III, Irving, CA USA 92697.

[9]University of Geneva Medical School, Department of Genetic Medicine and Development and iGE3 Institute of Genetics and Genomics of Geneva, 1 rue Michel-Servet, Geneva, Switzerland 1015.

[10]University of Lausanne, Center for Integrative Genomics, Genopode building, Lausanne, Switzerland 1015.

[11]University of North Carolina at Chapel Hill, Department of Biochemistry & Biophysics, 120 Mason Farm Rd., Chapel Hill, NC USA 27599.

[12]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire United Kingdom CB10 1SA.

[13]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520.

[14]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520.

[15]Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520.

[16]St. Laurent Institute, One Kendall Square, Cambridge, MA.

[17]Department of Chemical Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

Acknowledgements

This work was supported by the National Human Genome Research Institute (NHGRI) production grants number U54HG004557, U54HG004555, U54HG004576 and U54HG004558, and by the NHGRI pilot grant number R01HG003700. It was also supported by the NHGRI ARRA stimulus grant 1RC2HG005591, the National Science Foundation (SNF) grant number 127375, the European Research Council (ERC) grant number 249968, a research grant for the RIKEN Omics Science Center from the Japanese Ministry of Education, Culture, Sports, Science and technology, and grants BIO2011-26205, CSD2007-00050, and INB GNV-1 from the Spanish Ministry of Science. We would also like to thank Chris Gunter and Wendy Spitzer for editorial assistance with the manuscript.

References

1. Mattick JS. Long noncoding RNAs in cell and developmental biology. Semin Cell Dev Biol. 2011; 22:327. doi:S1084-9521(11)00077-2 [pii] 10.1016/j.semcdb.2011.05.002. [PubMed: 21621631]
2. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004; 306:636–640. doi: 306/5696/636 [pii] 10.1126/science.1105136. [PubMed: 15499007]
3. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. doi:10.1038/nature05874. [PubMed: 17571346]
4. Kapranov P, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007; 316:1484–1488. doi:1138341 [pii] 10.1126/science.1138341. [PubMed: 17510325]
5. Kapranov P, Willingham AT, Gingeras TR. Genome-wide transcription and the implications for genomic organization. Nat Rev Genet. 2007; 8:413–423. doi:nrg2083 [pii] 10.1038/nrg2083. [PubMed: 17486121]
6. Coffey AJ, et al. The GENCODE exome: sequencing the complete human exome. Eur J Hum Genet. 2011; 19:827–831. doi:ejhg201128 [pii] 10.1038/ejhg.2011.28. [PubMed: 21364695]
7. Harrow J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006; 7(Suppl 1):S4, 1–9. doi:gb-2006-7-s1-s4 [pii] 10.1186/gb-2006-7-s1-s4. [PubMed: 16925838]
8. Harrow, J. e. a. GENCODE: The reference human genome annotation for the ENCODE project. Genome research. 2012; XXX
9. Kodzius R, et al. CAGE: cap analysis of gene expression. Nat Methods. 2006; 3:211–222. doi:nmeth0306-211 [pii] 10.1038/nmeth0306-211. [PubMed: 16489339]
10. Ng P, et al. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods. 2005; 2:105–111. doi:nmeth733 [pii] 10.1038/nmeth733. [PubMed: 15782207]
11. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Annals of Applied Statistics. 2011; 5:1752–1779.
12. Cheng J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005; 308:1149–1154. doi:1108625 [pii] 10.1126/science.1108625. [PubMed: 15790807]
13. Katinakis PK, Slater A, Burdon RH. Non-polyadenylated mRNAs from eukaryotes. FEBS Lett. 1980; 116:1–7. doi:0014-5793(80)80515-1 [pii]. [PubMed: 6997068]
14. Milcarek C, Price R, Penman S. The metabolism of a poly(A) minus mRNA fraction in HeLa cells. Cell. 1974; 3:1–10. doi:0092-8674(74)90030-0 [pii]. [PubMed: 4213457]
15. Salditt-Georgieff M, Harpold MM, Wilson MC, Darnell JE Jr. Large heterogeneous nuclear ribonucleic acid has three times as many 5′ caps as polyadenylic acid segments, and most caps do not enter polyribosomes. Mol Cell Biol. 1981; 1:179–187. [PubMed: 6152852]
16. Khatun J, et al. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. Genome research. 2012; XXX
17. Tilgner H, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome research. 2012; XXX
18. Tilgner H, et al. Genomic analysis of ENCODE data reveals widespread links between epigenetic chromatin marks andalternative splicing. Genome research. 2012; XXX
19. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. doi:nmeth.1226 [pii] 10.1038/nmeth.1226. [PubMed: 18516045]
20. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature. 2009; 457:1028–1032. doi:nature07759 [pii] 10.1038/nature07759. [PubMed: 19169241]
21. consortium, T. E. p. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature. 2012; XXX
22. Thurman, R. E. e. a. The accessible chromatin landscape of the human genome. Nature. 2012; XXX

23. Gerstein, M. B. e. a. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012; XXX
24. Wang, J. e. a. Genome-wide mapping of the binding sites of 119 human transcription factors. Nature. 2012; XXX
25. Fu Y, et al. Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome research. 2011; 21
26. Ameur A, et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nature structural & molecular biology. 2011; 18:1435–1440. doi:10.1038/nsmb.2143.
27. Cole C, et al. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA. 2009; 15:2147–2160. doi:rna.1738409 [pii] 10.1261/rna.1738409. [PubMed: 19850906]
28. Kawaji H, et al. Hidden layers of human small RNAs. BMC Genomics. 2008; 9:157. doi:1471-2164-9-157 [pii] 10.1186/1471-2164-9-157. [PubMed: 18402656]
29. Lee YS, Shibata Y, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. 2009; 23:2639–2649. doi:23/22/2639 [pii] 10.1101/gad.1837609. [PubMed: 19933153]
30. Park E, Williams B, Wold B, Mortazavi A. A Survey of RNA Editing in the human ENCODE RNA-seq data (GRCP043). Genome research. 2012; XXX
31. Li M, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011; 333:53–58. [PubMed: 21596952]
32. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011; 7:522. doi:msb201154 [pii] 10.1038/msb.2011.54. [PubMed: 21811232]
33. Faulkner GJ, et al. The regulated retrotransposon transcriptome of mammalian cells. Nature genetics. 2009; 41:563–571. doi:ng.368 [pii] 10.1038/ng.368. [PubMed: 19377475]
34. Kim TK, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010; 465:182–187. doi:nature09033 [pii] 10.1038/nature09033. [PubMed: 20393465]
35. Ren B. Transcription: Enhancers make non-coding RNA. Nature. 2010; 465:173–174. doi:465173a [pii] 10.1038/465173a. [PubMed: 20463730]
36. Wang D, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature. 2011; 474:390–394. doi:nature10006 [pii] 10.1038/nature10006. [PubMed: 21572438]
37. Yip KY, et al. Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. Genome biology. 2012 (in press).
38. Hoffman, M. e. a. Integrative annotation of chromatin elements from encode data. Genome research. 2012; XXX
39. Arvey A, Agius P, Noble WS, Leslie C. Sequence and chromatin determinants of cell-type specific transcription factor binding. Genome research. 2012; XXX
40. Kundaje, A. e. a. Ubiquitous heterogeneity and asymmetry of the chromatin landscape at transcription regulatory elements. Genome research. 2012; XXX
41. Miller, B. e. a. Pre-programming of chromatin structure across the cell cycle. Genome research. 2012; XXX

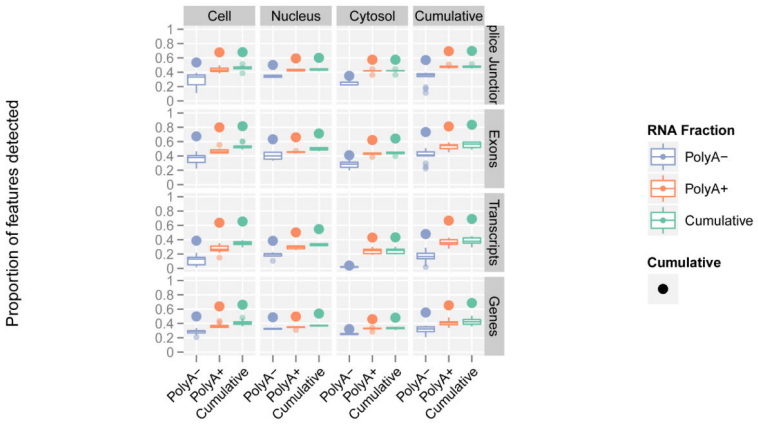

Figure1. A large majority of Gencode elements are detected by RNA-seq data

Shown are Gencode detected elements in the polyadenylated and non-polyadenylated fractions of cellular compartments (cumulative counts for both RNA fractions and compartments refer to elements present in any of the fractions or compartments). Each box plot is generated from values across all cell lines, thus capturing the dispersion across cell lines. The largest point shows the cumulative value over all cell lines.

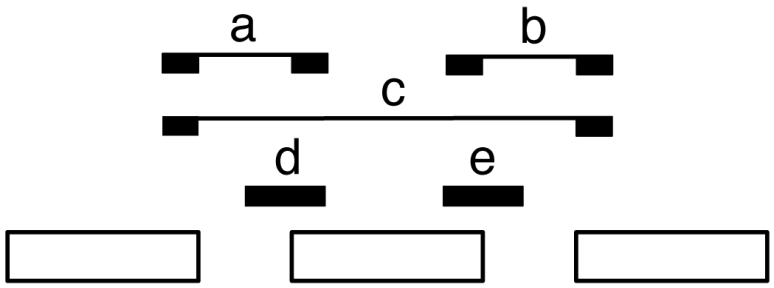

Figure2. Co-transcriptional splicing

a. Short read mappings for exon-based splicing completion. Read mappings that allow assessment of splicing completion around exons. (*a*,*b*,*c*) Reads providing evidence of splicing completion for the region containing the exon (with either exon inclusion, *ab*, or exclusion, *c*) (*d*,*e*) Reads providing evidence for the splicing of the region containing the exon not being completed yet. The complete Splicing Index (*coSI*) is the ratio of *a*+*b*+*c* over *a*+*b*+*c*+*d*+*e* and can thus be broadly assumed to correspond to the fraction of RNA molecules in which the region containing the exon has already been spliced (see Tilgner et al.[17]). A *coSI* value of 1 means splicing completed, while a value of 0 indicates that splicing has not yet been initiated.

b. Distribution of coSI scores computed on Gencode internal exons: (Top) Distribution in the total chromatin RNA fraction. (Bottom) Distribution in cytosolic polyadenylated RNA fraction.

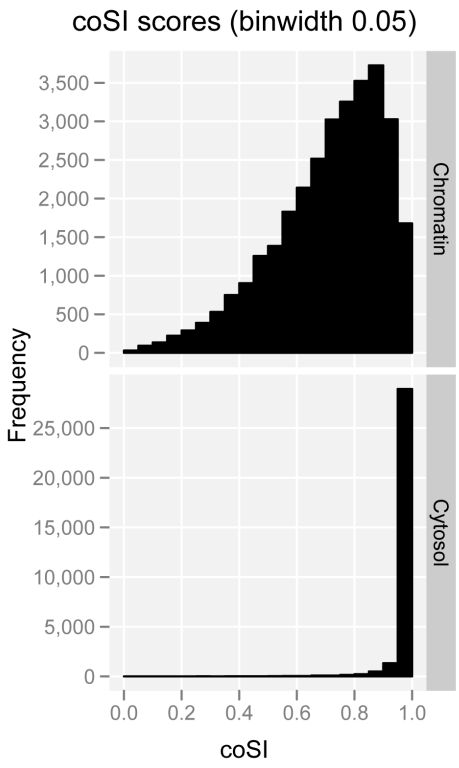

Figure 3. Abundance of gene types in cellular compartments

2D Kernel density plots of nuclear over cytosolic enrichment (Y axis) versus overall gene expression in the whole cell extract (X axis), for protein coding, long non-coding and novel genes over all cell lines. Only genes present in all 3 RNA extracts are displayed, as well as two representative genes (*ACTG1* in red and *H19* in blue), for which the expression in each individual cell line is shown. The actual values of the estimated Kernel density are indicated by contour lines and color shades.

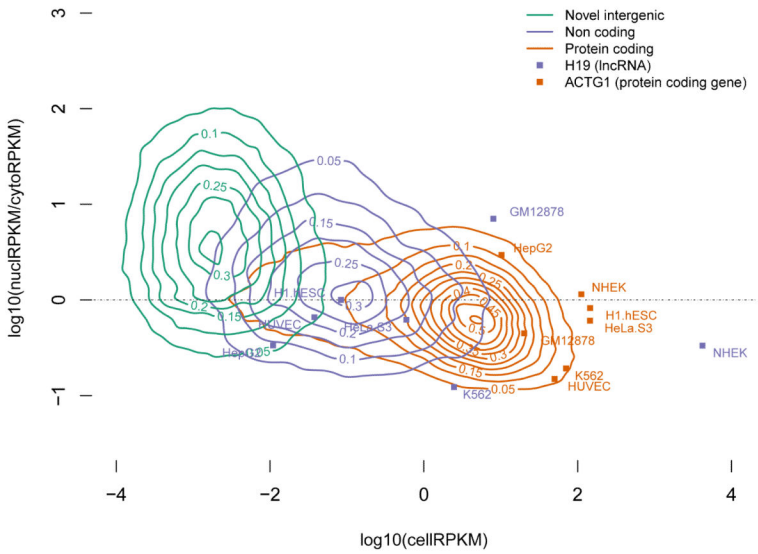

Figure 4. Isoform expression within a gene

a. Number of expressed isoforms per gene per cell line. Genes tends to express many isoforms simultaneously.

b. Relative expression of the most abundant isoform per gene per cell line. There is generally one dominant isoform in a given condition.

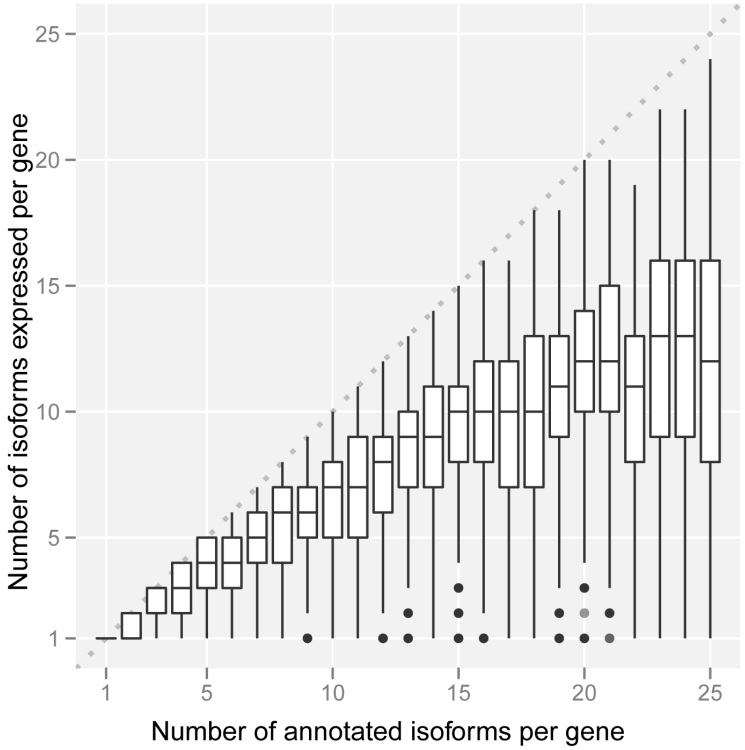

Figure 5. Transcription at enhancers

a. The pattern of RNA elements around enhancer predictions[21,37] containing DNase I hypersensitive (HS) sites. The lines represent the average frequency of RNA elements (top: polyadenylated long RNA contigs; middle: CAGE tag clusters; bottom: non-polyadenylated long RNA contigs) in a genomic window around the center of the enhancer prediction as determined by DNase I HS sites. Elements on the plus strand are shown in red, and on the minus strand in blue.

b. Enhancer transcripts differ from promoter transcripts.

The box plots compare the features of transcripts at predicted enhancer loci compared to predicted novel intergenic promoters[21] and annotated promoters[8]. H3k4me3, PolyA+ and Nucleus denote the 3 following ratios: H3k4me3/(H3k4me3 + H3k4me1), polyadenylated/(polyadenylated + non-polyadenylated), Nuclear/(Nuclear + Cytosolic). Enhancers are marked by higher levels of H3k4me1 compared to H3K4me3 than novel or annotated promoters (left). Enhancer transcripts show higher levels of non-polyadenylated (middle) and nuclear (right) RNA relative to promoters.

c. Chromatin state at transcribed enhancers.

Enhancer predictions with evidence of transcription (in blue; Cage tags present at predicted locus) show a different pattern of histone modifications and higher levels of RNA Polymerase II binding than non-transcribed predictions (red). They are enriched for H3K27 acetylation, H3K4 methylation, H3K79 di-methylation and depleted for H3K27 tri-methylation.

d. Enhancer activity and transcription is cell type specific.

Loci predicted to be active transcribed enhancers in GM12878 cells, show low signal for CAGE tags (top) and for H3K27 acetylation (bottom) in other cell lines.

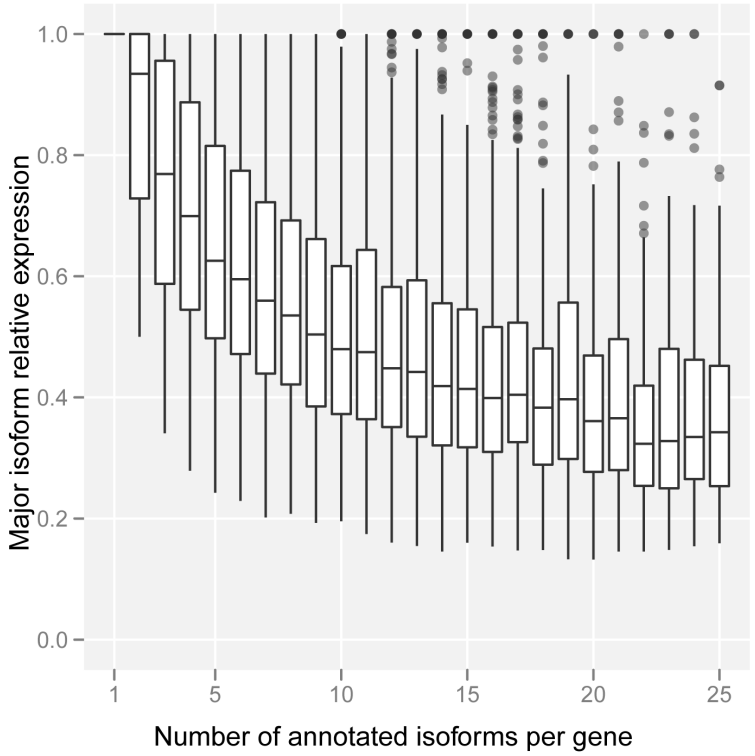

Figure 6. Size distribution of intergenic regions

Novel genes increase the proportion of small intergenic regions; ig/as = intergenic / antisense.

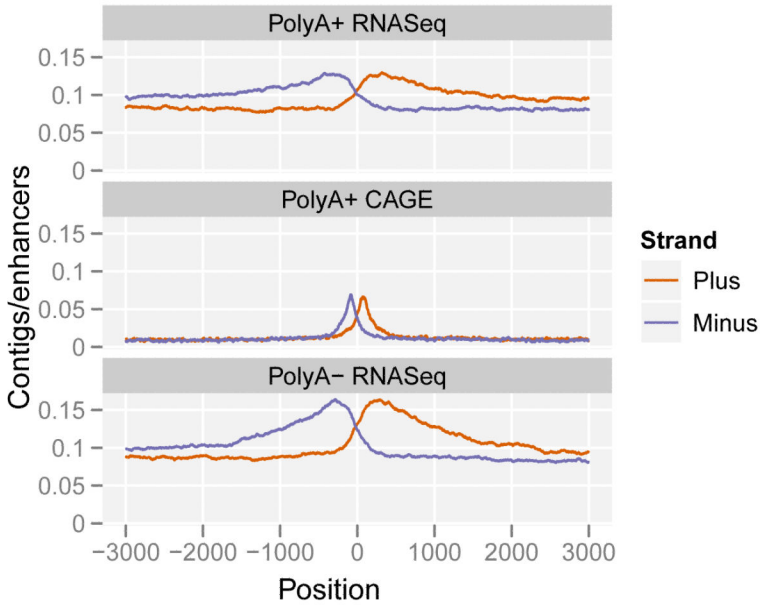

Figure 7.

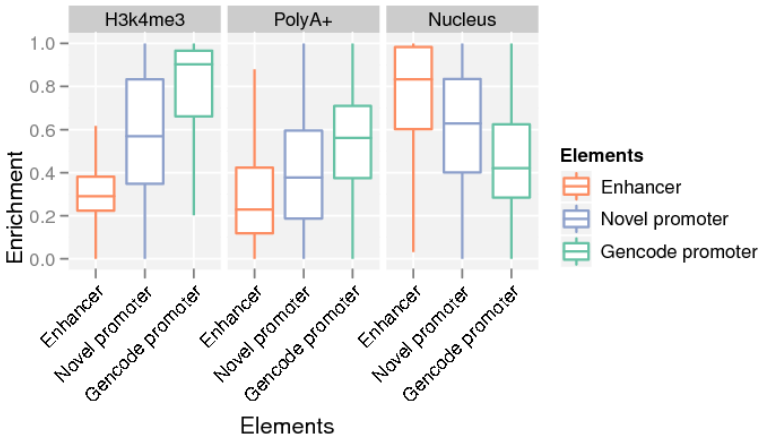

Figure 8.

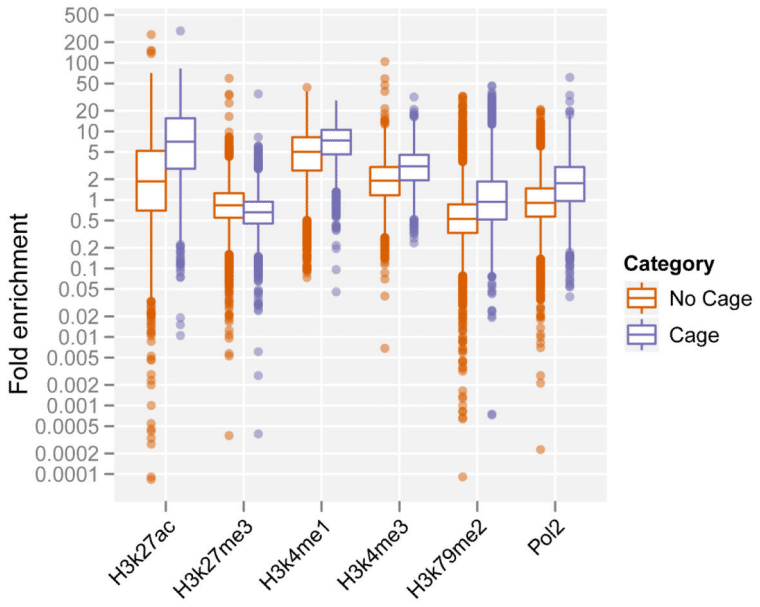

Figure 9.

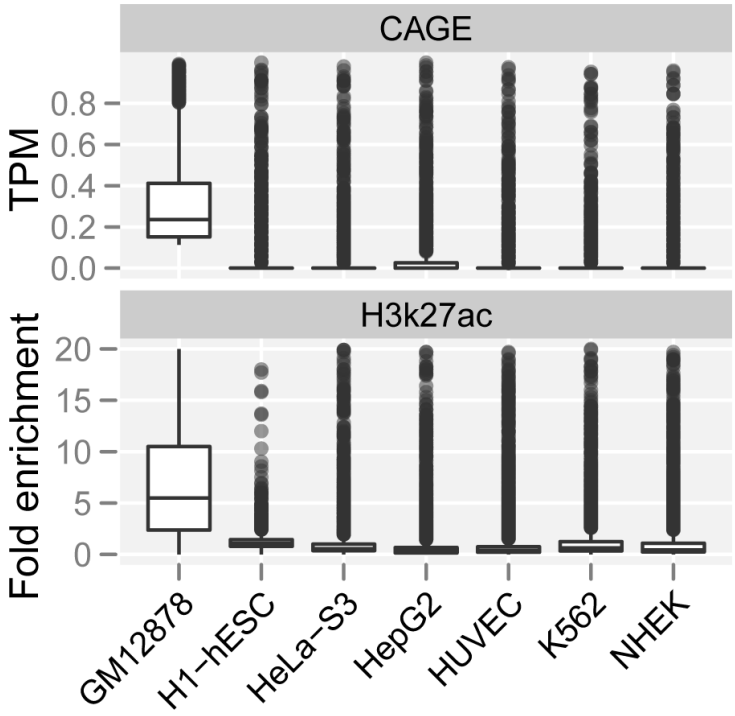

Figure 10.

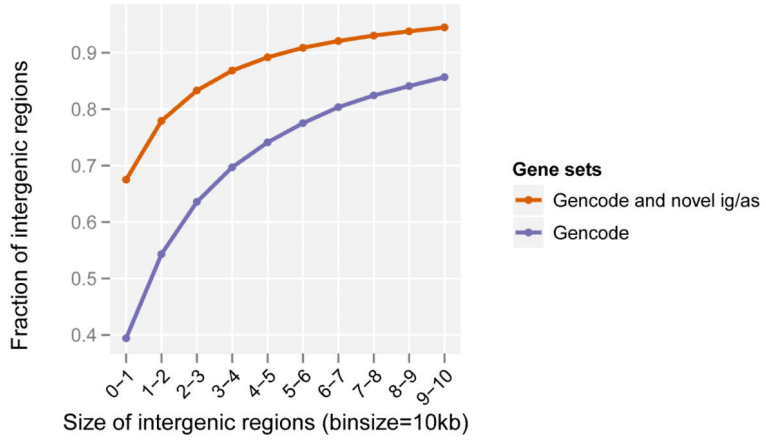

Figure 11.

Table 1

Long polyadenylated and non polyadenylated RNAs

1. Expression of Gencode (v7) annotated elements

Gene type	Detected exons[2] (annotation #)	Detected splice junctions[2] (annotation #)	Detected transcripts[2] (annotation #)	Detected genes[2] (annotation #)	Exon nucleotide coverage[3] (%)	Number of genes expressed in at least one cell line	Number of genes expressed in only 1 cell line	Proportion over genes expressed (%)	Number of genes expressed in 14 cell lines	Proportion over genes expressed (%)
Long non coding	22,381 (41,467)	8,017 (26,872)	6,521 (14,880)	5,906 (9,277)	87.5	5,906	1,386	23.5	631	10.7
Protein coding	288,322 (318,514)	194,752 (244,158)	59,822 (76,006)	18,939 (20,679)	98.1	18,939	1,082	5.7	10,571	55.8
Other[1]	102,000 (133,937)	19,277 (47,663)	45,410 (71,113)	10,649 (21,750)	95.2	10,649	2,453	23.0	1,896	17.8
Total annotated	412,703 (493,918)	222,046 (318,693)	111,753 (161,999)	35,494 (51,706)	96.7	35,394	4,921	13.9	13,098	37.0

2. Expression of Gencode (v7) intergenic and antisense elements

Category	Detected exons[2]	Detected splice junction[2]	Detected transcripts[2]	Detected genes[2]
Mono-exonic	55,683	NA	55,682	33,686
Multi-exonic	39,117	69,052	17,643	7,518
Total	94,800	69,052	73,325	41,204

[1] includes pseudogenes, miRNAs, etc

[2] all elements that passed npIDR (0.1)

[3] cumulative detected nucleotide in detected exons / total nucleotides in detected exons

Table 2

Short RNAs

a. Expression of Gencode (v7) annotated small RNA genes

Gene type[1].	Gencode total	Detected genes (% detected)	# Genes expressed in only 1 cell line (% detected)	# Genes expressed in 12 cell lines (% detected)	miRNA guide fragment[3]	miRNA passenger fragment[4]	Internal fragments[5] of annotated small RNA (average per detected gene)
miRNA	1,756	497 (28)	59 (12)	147 (30)	454 (454)	175 (175)	18
snoRNA	1,521	458 (30)	73 (16)	223 (49)	NA	NA	60
snRNA	1,944	378 (19)	123 (33)	41 (11)	NA	NA	36
tRNA	624	465 (75)	29 (6)	197 (42)	NA	NA	52
Other[2]	1,209	191 (16)	69 (36)	24 (13)	NA	NA	32
Total Gencode	7,054	1,989 (28)	353 (18)	632 (32)	NA	NA	40

b. Expression of unannotated short RNAs

Cell compartment	Unannotated short RNAs	Exonic	Intronic	Exon-intron boundaries	Genic	Gene-intergene boundaries	Intergenic
cell	57,393	14,116	13,773	1,818	29,707	13,048	25,906
nucleus	82,297	19,334	40,136	5,248	64,718	7,417	16,289
cytosol	25,455	6,183	5,605	665	12,453	6,631	12,447
3 compartments	150,165	38,969	55,061	7,552	101,582	23,185	45,081

[1] includes all other Gencode small transcripts biotypes except pseudogenes

[2] all elements that have passed npIDR (0.1)

[3] number of detected miRNAs with an expressed annotated guide (with an annotated guide in mirbase)

[4] number of detected miRNAs with an expressed annotated passenger (with an annotated passenger in mirbase)

[5] short RNAseq mapping which 5′ ends starts 5 bp after the start and ends 5bp before the end of a detected gene