ARTICLE

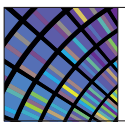
doi:10.1038/nature11233

Landscape of transcription in human cells

Sarah Djebali[1]*, Carrie A. Davis[2]*, Angelika Merkel[1], Alex Dobin[2], Timo Lassmann[3], Ali Mortazavi[4,5], Andrea Tanzer[1], Julien Lagarde[1], Wei Lin[2], Felix Schlesinger[2], Chenghai Xue[2], Georgi K. Marinov[4], Jainab Khatun[6], Brian A. Williams[4], Chris Zaleski[2], Joel Rozowsky[7,8], Maik Röder[1], Felix Kokocinski[9], Rehab F. Abdelhamid[3], Tyler Alioto[1,10], Igor Antoshechkin[4], Michael T. Baer[2], Nadav S. Bar[11], Philippe Batut[2], Kimberly Bell[2], Ian Bell[12], Sudipto Chakrabortty[2], Xian Chen[13], Jacqueline Chrast[14], Joao Curado[1], Thomas Derrien[1], Jorg Drenkow[2], Erica Dumais[12], Jacqueline Dumais[12], Radha Duttagupta[12], Emilie Falconnet[15], Meagan Fastuca[2], Kata Fejes-Toth[2], Pedro Ferreira[1], Sylvain Foissac[12], Melissa J. Fullwood[16], Hui Gao[12], David Gonzalez[1], Assaf Gordon[2], Harsha Gunawardena[13], Cedric Howald[14], Sonali Jha[2], Rory Johnson[1], Philipp Kapranov[12,17], Brandon King[4], Colin Kingswood[1,10], Oscar J. Luo[16], Eddie Park[5], Kimberly Persaud[2], Jonathan B. Preall[2], Paolo Ribeca[1,10], Brian Risk[6], Daniel Robyr[15], Michael Sammeth[1,10], Lorian Schaffer[4], Lei-Hoon See[2], Atif Shahab[16], Jorgen Skancke[1,11], Ana Maria Suzuki[3], Hazuki Takahashi[3], Hagen Tilgner[1]†, Diane Trout[4], Nathalie Walters[14], Huaien Wang[2], John Wrobel[6], Yanbao Yu[13], Xiaoan Ruan[16], Yoshihide Hayashizaki[3], Jennifer Harrow[9], Mark Gerstein[7,8,18], Tim Hubbard[9], Alexandre Reymond[14], Stylianos E. Antonarakis[15], Gregory Hannon[2], Morgan C. Giddings[6,13], Yijun Ruan[16], Barbara Wold[4], Piero Carninci[3], Roderic Guigó[1,19] & Thomas R. Gingeras[2,12]

Eukaryotic cells make many types of primary and processed RNAs that are found either in specific subcellular compartments or throughout the cells. A complete catalogue of these RNAs is not yet available and their characteristic subcellular localizations are also poorly understood. Because 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, modification 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 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 functions carried out 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 five-year 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 previous studies[4,5]. During the second phase of the ENCODE project, lasting 5 years, the scope of examination was broadened to interrogate the complete human genome. Thus, we have sought to both provide a genome-wide catalogue of human transcripts and to identify the subcellular 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 subcompartments (nucleus and cytosol) for all 15 cell lines studied, and in three additional subnuclear compartments in one cell line. In addition, we have sought to determine whether 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.



ENCODE
Encyclopedia of DNA Elements
nature.com/encode

- 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

[1]Centre for Genomic Regulation and UPF, Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain. [2]Cold Spring Harbor Laboratory, Functional Genomics, 1 Bungtown Road, Cold Spring Harbor, New York 11742, USA. [3]RIKEN Yokohama Institute, RIKEN Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [4]California Institute of Technology, Division of Biology, 2 Beckman Institute, Pasadena, California 91125, USA. [5]University of California Irvine, Department of Developmental and Cell Biology, 2300 Biological Sciences III, Irvine, California 92697, USA. [6]Boise State University, College of Arts & Sciences, 1910 University Drive, Boise, Idaho 83725, USA. [7]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [8]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [9]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [10]Centro Nacional de Análisis Genómico (CNAG), C/ Baldiri Reixac 4, Torre I, Barcelona 08028, Catalonia, Spain. [11]Department of Chemical Engineering, Norwegian University of Science and Technology, Trondheim NO-7491, Norway. [12]Affymetrix, Inc, 3380 Central Expressway, Santa Clara, California 95051, USA. [13]University of North Carolina at Chapel Hill, Department of Biochemistry & Biophysics, 120 Mason Farm Road, Chapel Hill, North Carolina 27599, USA. [14]University of Lausanne, Center for Integrative Genomics, Genopode building, Lausanne 1015, Switzerland. [15]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 1211, Switzerland. [16]Genome Institute of Singapore, Genome Technology and Biology, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [17]St Laurent Institute, One Kendall Square, Cambridge, Massachusetts 02141, USA. [18]Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [19]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain. †Present address: Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.
*These authors contributed equally to this work.

overlapping of neighbouring gene regions and prompts a redefinition of a gene.

- Isoform expression by a 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 I 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 probably precursors to these small RNAs. The subcellular localization of both annotated and unannotated short RNAs is highly specific.

RNA data set generation

We performed subcellular compartment fractionation (whole cell, nucleus and cytosol) before RNA isolation in 15 cell lines (Supplementary Table 1) to interrogate deeply the human transcriptome. For the K562 cell line, we also performed additional nuclear subfractionation into chromatin, nucleoplasm and nucleoli. The RNAs from each of these subcompartments were prepared in replica and were separated based on length into >200 nucleotides (long) and <200 nucleotides (short). Long RNAs were further fractionated into polyadenylated and non-polyadenylated transcripts. A number of complementary technologies were used 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]; Supplementary Fig. 1). Sequence reads were mapped and post-processed using a variety of software tools (Supplementary Table 2 and Supplementary Fig. 2). We used the mapped data to assemble and quantify *de novo* elements (exons, transcripts, genes, contigs, splice junctions and transcription start sites (TSSs)) 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 Information) 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 (DCC, http://genome.ucsc.edu/ENCODE/dataSummary.html) (Supplementary Fig. 2). 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 (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) 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

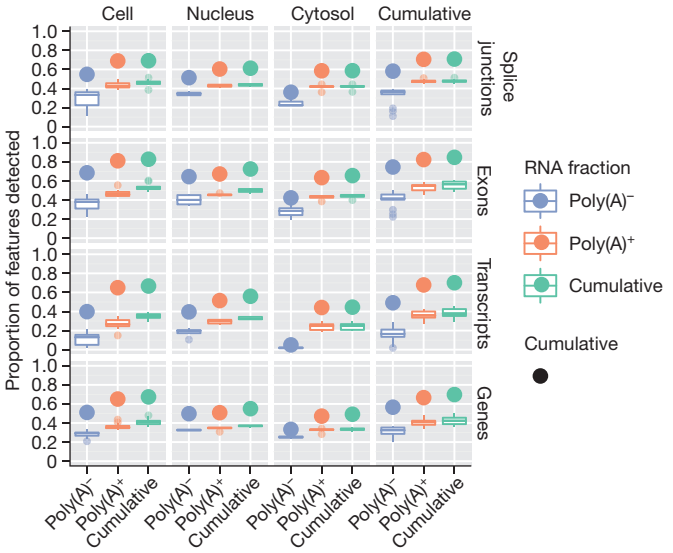


Figure 1 | A large majority of GENCODE elements are detected by RNA-seq data. Shown are GENCODE-detected elements in the polyadenylated and non-polyadenylated fractions of cellular compartments (cumulative counts for both RNA fractions and compartments refer to elements present in any of the fractions 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.

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 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. 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; J. Khatun *et al.*, manuscript in preparation), 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).

The interrogation of different subcellular RNA fractions provides snapshots of the status of the RNA population along the RNA processing pathway. Thus, by analysing short and long RNAs in the different subcellular compartments, we confirm that splicing predominantly occurs during transcription. By using RNA-seq to measure the degree of completion of splicing (Fig. 2a), we observed that around most exons, introns are already being spliced in chromatin-associated RNA—the fraction that includes RNAs in the process of being transcribed (Fig. 2b). Concomitantly, we found strong enrichment specifically of spliceosomal small nuclear RNAs (snRNAs) in this RNA fraction (see 'Short RNA expression landscape' later). Co-transcriptional splicing provides an explanation for the increasing evidence connecting chromatin structure to splicing regulation, and we have observed that exons in the process of being spliced are enriched in a number of chromatin marks[16,17].

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[18], 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).

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

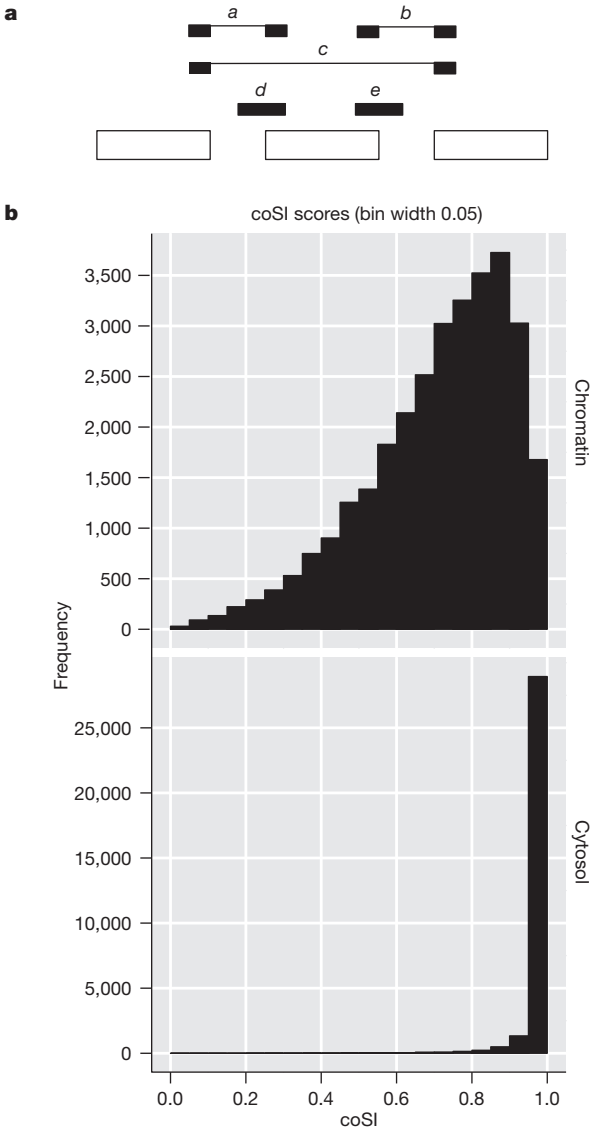


Figure 2 | Co-transcriptional splicing. a, Short read mappings for exon-based splicing completion. Read mappings that allow assessment of splicing completion around exons are shown. Reads providing evidence of splicing completion for the region containing the exon (with either exon inclusion (*a*, *b*) or exclusion (*c*)) are shown. Reads providing evidence for the splicing of the region containing the exon not being completed yet are indicated by *d* and *e*. The complete splicing index (coSI) is the ratio of $(0.5(a + b) + c)$ over $(0.5(a + b) + c + 0.5(d + e))$ 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 ref. 16). A coSI value of 1 means splicing completed, whereas 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.

cell lines studied when considering the whole cell polyadenylated RNAs, whereas only 10% were expressed in all cell lines. Conversely, whereas a large fraction (53%) of expressed protein-coding genes were constitutive (expressed in all cell lines), only ~7% were cell-line specific (Supplementary Table 7 and Supplementary Fig. 10).

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 (Fig. 4a). The increase, however, is not linear and seems to plateau at about 10–12 expressed

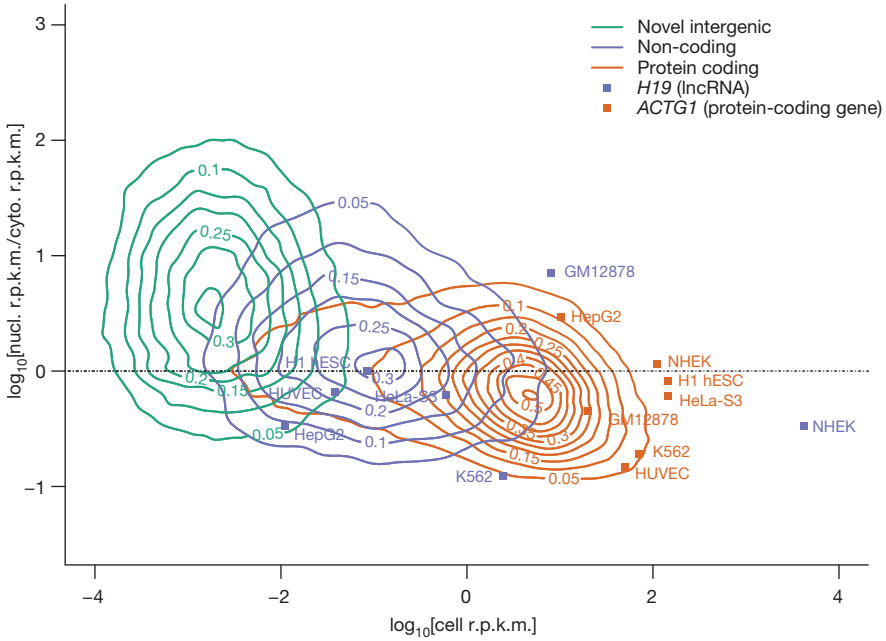


Figure 3 | Abundance of gene types in cellular compartments. Two-dimensional kernel density plots of nuclear over cytosolic enrichment (*y* axis) versus overall gene expression in the whole cell extract (*x* axis), for protein coding, long non-coding and novel genes over all cell lines. Only genes present in all three 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 colour shades.

isoforms per gene. However, we cannot obviously distinguish 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; Fig. 4b). Third, about three-quarters of protein-coding genes have at least two different dominant/major isoforms depending on the cell line (Supplementary Fig. 11a). 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 $1/n$ (Supplementary Fig. 11b). Fifth, variability of gene expression contributes more than variability of splicing ratios to the variability of transcript abundances across cell lines (Supplementary Information).

Alternative transcription initiation and termination

On the basis of RNA-seq analysis of polyadenylated RNAs, a total of 128,021 TSSs were detected across all cell lines, of which 97,778 were previously annotated and 30,243 were novel intergenic/antisense TSSs (Supplementary Table 3a). CAGE tags, filtered by a hidden Markov model (HMM)-based algorithm to differentiate between 5′ capped termini of polymerase II transcripts and recapping events[19] (Supplementary Information), identified a total of 82,783 non-redundant TSSs (Supplementary Table 8). Approximately 48% of the CAGE-identified TSSs are located within 500 base pairs (bp) of an annotated RNA-seq-detected GENCODE TSS, whereas an additional 3% are within 500 bp of a novel TSS (Supplementary Fig. 12). Notably, only ~72% of all CAGE sequencing reads map to TSSs, indicating that the remaining 30% may originate from recapping events or from a new class of TSS.

Using data collected within the ENCODE consortium[20], 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[21], chromatin modification and DNA binding elements[22,23]. All GENCODE/RNA-seq-determined TSSs were examined in each of the cell lines (Supplementary Fig. 13, column 1). Of these redundant positions, 44.7% (199,146) of the RNA-seq-supported TSSs also displayed

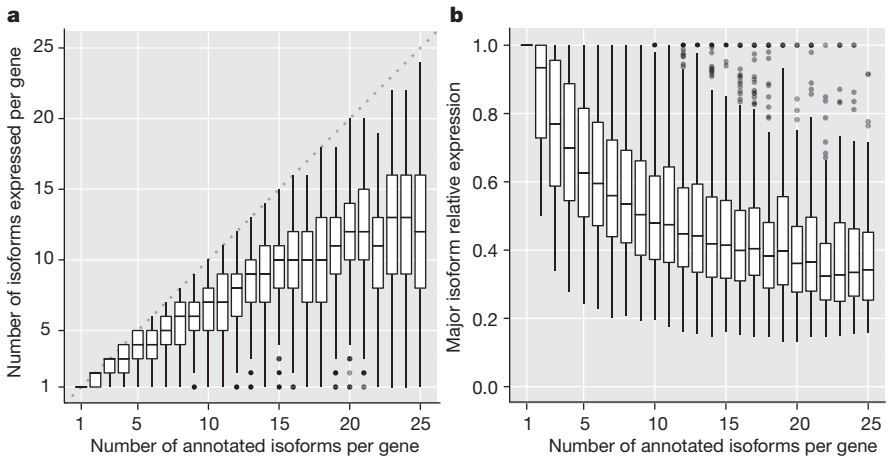


Figure 4 | Isoform expression within a gene. **a**, Number of expressed isoforms per gene per cell line. Genes tend to express many isoforms simultaneously. **b**, Relative expression of the most abundant isoform per gene per cell line. There is generally one dominant isoform in a given condition. The whiskers are defined as Q1 $-1.5\times$IQR to Q3 $+1.5\times$IQR, where IQR is the interquartile range, and Q1 and Q3 the first and third quartile, respectively. Each box plot was constructed using the number of genes with 1, 2, 3, 4, etc. up to 25 isoforms.

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 TSSs identified by either CAGE or RNA-seq/GENCODE displayed all of the characteristics of the start of transcription (presence of DNase I, H3K4me3, H3K27ac sites and either TAF1 or TBP binding). This is consistent with the possibility that regulatory regions proximal to TSSs are of more than one type.

At the 3′ end, 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[24]. About 20% of these mapped proximal to annotated polyadenylation sites (PAS) whereas 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 Information).

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 (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[16,25].

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[26–28].

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].

Genealogy of short RNAs

Genome wide, 27% of annotated small RNAs reside within 8% of protein-coding and 5% within 3% of lncRNA genes (Supplementary Fig. 16). Overall, about 6% of all annotated long transcripts overlap with small RNAs and are probably precursors to these small RNAs. Although most 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 (Supplementary Fig. 17a). Additionally, 8.4% of GENCODE annotated small RNAs map within novel intergenic transcripts, with most overlapping annotated tRNAs. The enrichment for tRNAs was mostly in novel intergenic transcripts derived from non-polyadenylated RNAs (Supplementary Fig. 17b). Many long RNAs, both novel and annotated, thus seem to have dual roles, as functional (protein coding) RNAs, and as precursors for many important classes of small RNAs. Using RNA-seq data from the K562 cell line, we investigated the preferential cellular localization of these RNA precursors (Supplementary Fig. 18). 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 (Supplementary Fig. 18a, d). Notably, whereas mature snRNAs were both nuclear and cytosolic, the overlapping long RNAs were observed to be primarily nuclear (Supplementary Fig. 18c). Finally, for snoRNAs (nuclear enrichment), potential long RNA precursors were decidedly observed to be both nuclear and cytosolic (Supplementary Fig. 18b). Unannotated short RNAs were found overall not to be enriched in either the nuclear or cytosolic compartment (Supplementary Fig. 18e).

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 artefacts and identify genes that are RNA edited[29]. Focusing first on GM12878, a cell line that has been deeply re-sequenced, we find a total 51,557 RNA consistent single nucleotide variants (SNVs) within genic boundaries, 65% of which are present in dbSNP. Of the remainder, 1,186 SNVs in 430 genes (Supplementary Fig. 19a) 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 is for T>C (5%) and these occur primarily in regions with detectable antisense transcription[29]. We find similar A>G(I) frequencies of 75–84% in seven additional cell lines (Supplementary Fig. 19b). 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[30].

Using the AlleleSeq pipeline[31] 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. The proportion of genes with allele-specific expression was similar in the three investigated RNA fractions (whole-cell, cytoplasm and nucleus; Supplementary Table 9 and Supplementary Information).

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 long interspersed element (LINE), short interspersed element (SINE), long terminal repeat (LTR) and other repeat elements, respectively (see Supplementary Information). Measuring Shannon entropy across cell lines, we found that CAGE clusters mapping to repeat regions were noticeably more narrowly expressed than CAGE clusters mapping within genic regions (Supplementary Fig. 20a). 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) (Supplementary Fig. 20b–d). Although a large proportion of the transcripts in the human genome is thought to be initiated from repetitive elements (especially retrotransposon elements[32]), 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[33–35]. We used our RNA assays to detect and characterize transcriptional activity at enhancer loci predicted genome-wide from ENCODE chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) data[20,36].

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 centred on those sites. In these plots, as denoted by the accumulation of CAGE tags signifying TSSs, transcription initiation within the enhancer region is observed, and continues outwards for several kilobases (kb). This behaviour can be observed for the polyadenylated and non-polyadenylated RNA fractions mapping in both intronic and intergenic regions. As previously reported[33], we observe a large diversity of expression levels at each of the transcribed enhancers. Polyadenylated to non-polyadenylated RNA ratios, as well as nuclear to cytoplasmic ratios, vary at individual enhancers (Supplementary Fig. 21a, b). However, contrary to some previous reports, although most eRNAs are prevalent in the nuclear non-polyadenylated RNA fraction, some eRNAs seemed to be polyadenylated in the nucleus. This pattern was significantly different compared to transcripts from GENCODE annotated and novel predicted[20] promoters (Fig. 5b).

Transcribed enhancers on average show a significantly different pattern of chromatin modification than non-transcribed ones[37–40]. 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

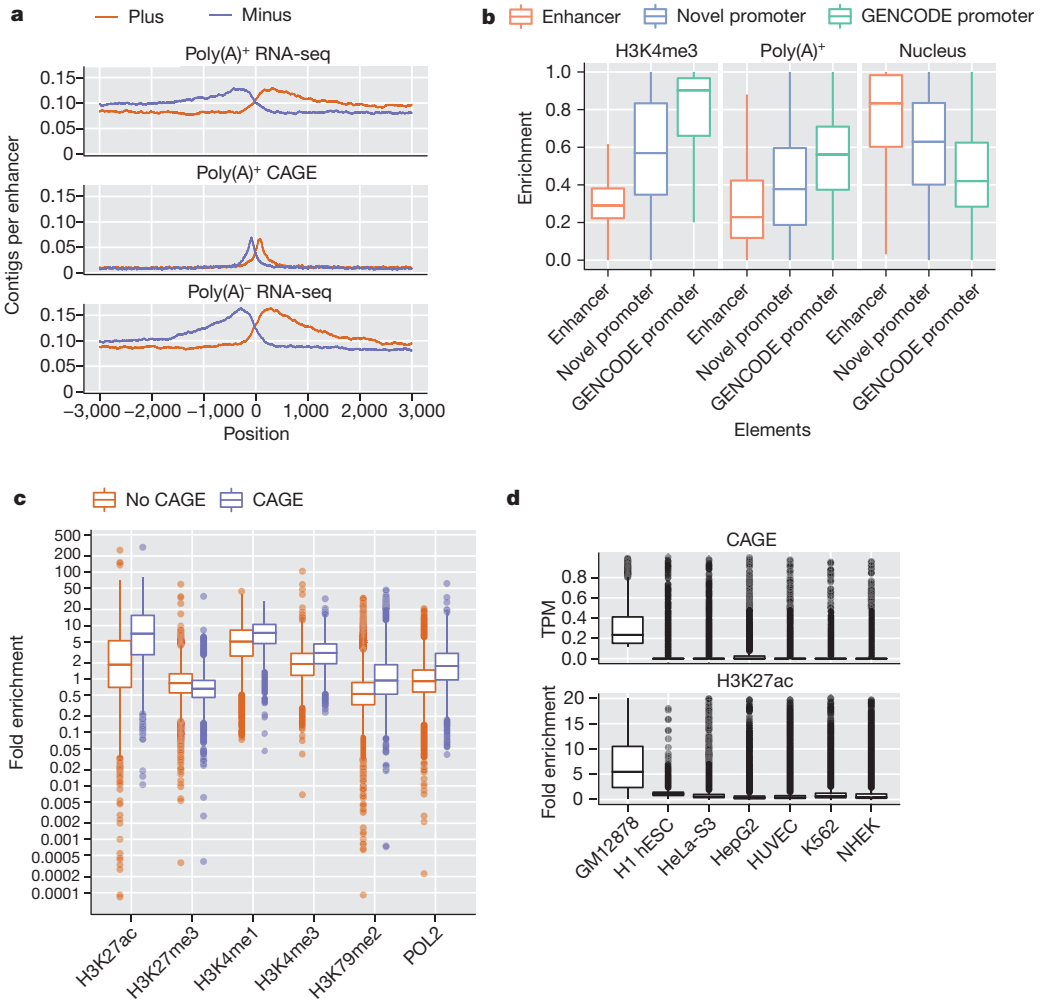


Figure 5 | Transcription at enhancers. **a**, The pattern of RNA elements around enhancer predictions[20,36] containing DNase I hypersensitive sites. The lines represent the average frequency of RNA elements (top, polyadenylated long RNA contigs; middle, CAGE tag clusters; bottom, non-polyadenylated long RNA contigs) in a genomic window around the centre of the enhancer prediction as determined by DNase I hypersensitive 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[20] and annotated promoters[8]. H3K4me3, poly(A)$^+$ and nucleus denote the three 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 modification and higher levels of RNA polymerase II binding than non-transcribed predictions (red). They are enriched for H3K27 acetylation, H3K4 methylation, H3K79 dimethylation and depleted for H3K27 trimethylation. **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. The whiskers are defined as Q1 $-1.5 \times$ IQR to Q3 $+1.5 \times$ IQR, where IQR is the interquartile range, and Q1 and Q3 the first and third quartile, respectively.

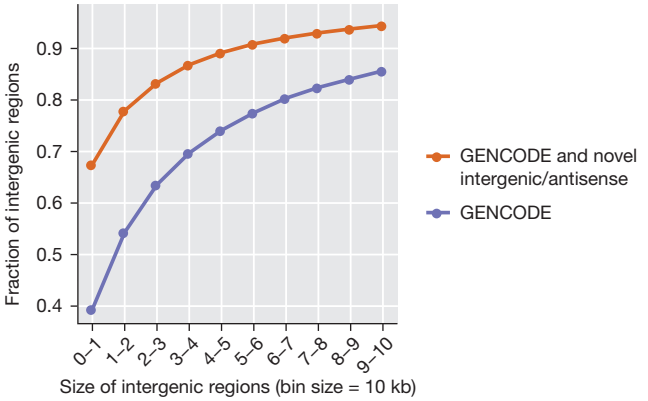


Figure 6 | Size distribution of intergenic regions. Novel genes increase the proportion of small intergenic regions.

transcriptional initiation and elongation (Fig. 5c). Both the transcripts and the chromatin states are cell-type specific (Fig. 5d). Taking the GM12878 cell line as an example, the enhancer loci producing eRNA demonstrate enrichment of CAGE tag detection (Fig. 5d, top) and the presence of H3K27ac histone modification (Fig. 5d, bottom) in this cell line compared to five other analysed 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.

Concluding remarks

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, respectively (Supplementary Table 10 and Supplementary Fig. 22). 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 (Supplementary Table 10), 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 (Supplementary Table 2.4.3 and ref. 3). The increased genome coverage by processed RNAs stems largely from the inclusion of

Table 1 | Long polyadenylated and non-polyadenylated RNAs

Expression of GENCODE (v7) annotated elements (a)

Gene type	Detected exons† (annotation no.)	Detected splice junctions† (annotation no.)	Detected transcripts† (annotation no.)	Detected genes† (annotation no.)	Exon nucleotide coverage‡ (%)	Number of genes expressed in at least one cell line	Number of genes expressed in only one 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*	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

Expression of GENCODE (v7) intergenic and antisense elements (b)

Category	Detected exons†	Detected splice junction†	Detected transcripts†	Detected genes†
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

NA, not applicable.
* Includes pseudogenes, miRNAs, etc.
† All elements that passed npIDR (0.1).
‡ Cumulative detected nucleotide in detected exons/total nucleotides in detected exons.
§ Proportion for genes expressed in only one cell line.
‖ Proportion for genes expressed in 14 cell lines.

Table 2 | Short RNAs

Expression of GENCODE (v7) annotated small RNA genes (a)

Gene type*	GENCODE total	Detected genes (% detected)	No. genes expressed in only one cell line (% detected)	No. genes expressed in 12 cell lines (% detected)	miRNA guide fragment‡	miRNA passenger fragment§	Internal fragments‖ 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†	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

Expression of unannotated short RNAs (b)

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
Three compartments	150,165	38,969	55,061	7,552	101,582	23,185	45,081

NA, not applicable.
* Includes all other GENCODE small transcript biotypes except for pseudogenes.
† All elements that have passed npIDR (0.1).
‡ Number of detected miRNAs with an expressed annotated guide (with an annotated guide in mirbase).
§ Number of detected miRNAs with an expressed annotated passenger (with an annotated passenger in mirbase).
‖ Short RNA-seq mapping for which the 5′ end starts 5 bp after the start and ends 5 bp before the end of a detected gene.

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 14,170 bp to 3,949 bp median length; Fig. 6). Concordantly, we observed an increased overlap of genic regions. As 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. As this is 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. Co-published ENCODE-related papers can be explored online via the *Nature* ENCODE explorer (http://www.nature.com/ENCODE), a specially designed visualization tool that allows users to access the linked papers and investigate topics that are discussed in multiple papers via thematically organized threads.

METHODS SUMMARY

For full details of Methods, see Supplementary Information.

Received 10 December 2011; accepted 15 May 2012.

1. Mattick, J. S. Long noncoding RNAs in cell and developmental biology. *Semin. Cell Dev. Biol.* **22,** 327 (2011).
2. The ENCODE Project Consortium.. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
3. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
4. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
5. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8,** 413–423 (2007).
6. Coffey, A. J. *et al.* The GENCODE exome: sequencing the complete human exome. *Eur. J. Hum. Genet.* **19,** 827–831 (2011).
7. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (suppl. 1), 1–9 (2006).
8. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* (in the press).
9. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature Methods* **3,** 211–222 (2006).
10. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods* **2,** 105–111 (2005).
11. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).
12. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308,** 1149–1154 (2005).
13. Katinakis, P. K., Slater, A. & Burdon, R. H. Non-polyadenylated mRNAs from eukaryotes. *FEBS Lett.* **116,** 1–7 (1980).
14. Milcarek, C., Price, R. & Penman, S. The metabolism of a poly(A) minus mRNA fraction in HeLa cells. *Cell* **3,** 1–10 (1974).
15. Salditt-Georgieff, M., Harpold, M. M., Wilson, M. C. & Darnell, J. E. 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.* **1,** 179–187 (1981).
16. 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 Res.* (in the press).
17. Tilgner, H. *et al.* Genomic analysis of ENCODE data reveals widespread links between epigenetic chromatin marks and alternative splicing. *Genome Res.* (in the press).
18. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
19. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457,** 1028–1032 (2009).
20. ENCODE Project Consortium.. An integrated encyclopaedia of DNA elements in the human genome. *Nature* http://dx.doi.org/10.1038/nature11247 (this issue).
21. Thurman, R. E. The accessible chromatin landscape of the human genome. *Nature* http://dx.doi.org/10.1038/nature11232 (this issue).
22. Gerstein, M. B. Architecture of the human regulatory network derived from ENCODE data. *Nature* http://dx.doi.org/10.1038/nature11245 (this issue).
23. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* (in the press).
24. Fu, Y. *et al.* Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* **21,** 741–747 (2011).
25. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Struct. Mol. Biol.* **18,** 1435–1440 (2011).
26. Cole, C. *et al.* Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* **15,** 2147–2160 (2009).
27. Kawaji, H. *et al.* Hidden layers of human small RNAs. *BMC Genom.* **9,** 157 (2008).
28. Lee, Y. S., Shibata, Y., Malhotra, A. & Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* **23,** 2639–2649 (2009).
29. Park, E., Williams, B., Wold, B. & Mortazavi, A. A Survey of RNA Editing in the human ENCODE RNA-seq data (GRCP043). *Genome Res.* (in the press).
30. Li, M. *et al.* Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333,** 53–58 (2011).
31. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
32. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nature Genet.* **41,** 563–571 (2009).
33. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
34. Ren, B. Transcription: Enhancers make non-coding RNA. *Nature* **465,** 173–174 (2010).
35. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474,** 390–394 (2011).
36. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. *Genome Biol.* (in the press).
37. Hoffman, M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Genome Res.* (in the press).
38. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type specific transcription factor binding. *Genome Res.* (in the press).
39. Kundaje, A. Ubiquitous heterogeneity and asymmetry of the chromatin landscape at transcription regulatory elements. *Genome Res.* (in the press).
40. Miller, B. Pre-programming of chromatin structure across the cell cycle. *Genome Res.* (in the press).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Human Genome Research Institute (NHGRI) production grants U54HG004557, U54HG004555, U54HG004576 and U54HG004558, and by the NHGRI pilot grant R01HG003700. It was also supported by the NHGRI ARRA stimulus grant 1RC2HG005591, the National Science Foundation (SNF) grant 127375, the European Research Council (ERC) grant 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 C. Gunter and W. Spitzer for editorial assistance with the manuscript.

Author Contributions 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. and Y.H. led the project and oversaw the analysis. C.A.D., X.R., B.A.W. and P.C. oversaw or significantly contributed to data generation. S.D., A.Me., A.D., T.L., A.Mo., A.T., J.L., W.L., F.S., C.X., G.K.M., J.K., C.Z., J.R., M.R., F.K. and J.H. made major contributions towards data processing 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.Ch., J.Cu., T.D., J.Dr., E.D., J.Du., R.D., E.F., M.F., K.F-T., P.F., S.F., M.J.F., H.Ga., D.G., A.G., H.Gu., 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., L.-H.S., A.S., J.S., A.M.S., H.Ta., H.Ti., D.T., N.W., H.W., J.W. and Y.Y. were responsible for data production and analysis. T.R.G. and R.G. wrote the manuscript with input from all authors.

Author Information A complete set of data files can be downloaded at GEO under accession codes GSE26284 (CSHL, long RNA), GSE33480 (Caltech, A+ RNA-seq), GSE24565 (CSHL, short RNA), GSE33600 (GIS, RNA-PET) and GSE34448 (RIKEN, CAGE) or are viewable at the UCSC Genome Browser (http://genome-preview.ucsc.edu/ENCODE/). Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R.G. (gingeras@cshl.edu) or R.G. (roderic.guigo@crg.eu).