3 Characterization of intergenic regions and gene definition

The prevalence and analysis of ENCODE data are changing the definition and characterization of intergenic and genic regions

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 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. 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. 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.

The annotation of all pseudogenes in the human reference genome is part of the wider effort by the GENCODE consortium which also aims to identify all protein-coding, long non-coding RNA and short RNA genes[27, 28]. Similar to the annotation of other functional classes, the annotation of pseudogenes contains models that have been created by the Human And Vertebrate Analysis aNd Annotation team (HAVANA), an expert manual annotation team at the Wellcome Trust Sanger Institute. This is informed by, and checked against, computational pseudogene predictions by the PseudoPipe[35] and Retrofinder[36] pipelines (details in Methods). These computational pseudogene predictions provide hints to manual annotators during the first-pass of annotation and identify potential missing features, flagging them for manual re-investigation (Fig. 1).

In this study, we focused on a pseudogene set comprised of manually annotated pseudogenes (a union of levels 1 and 2). Polymorphic pseudogenes, which are coding genes that are pseudogenic due to the presence of a polymorphic premature stop codon in the reference genome (GRCh37), were excluded from our study in order to avoid the likelihood that they may have coding potential in the cell lines and tissues studied by other ENCODE groups. We call these 11,216 pseudogenes the "surveyed set". The set contains 138 unitary pseudogenes. For the purpose of this paper, only the processed and duplicated pseudogenes will be discussed in detail.

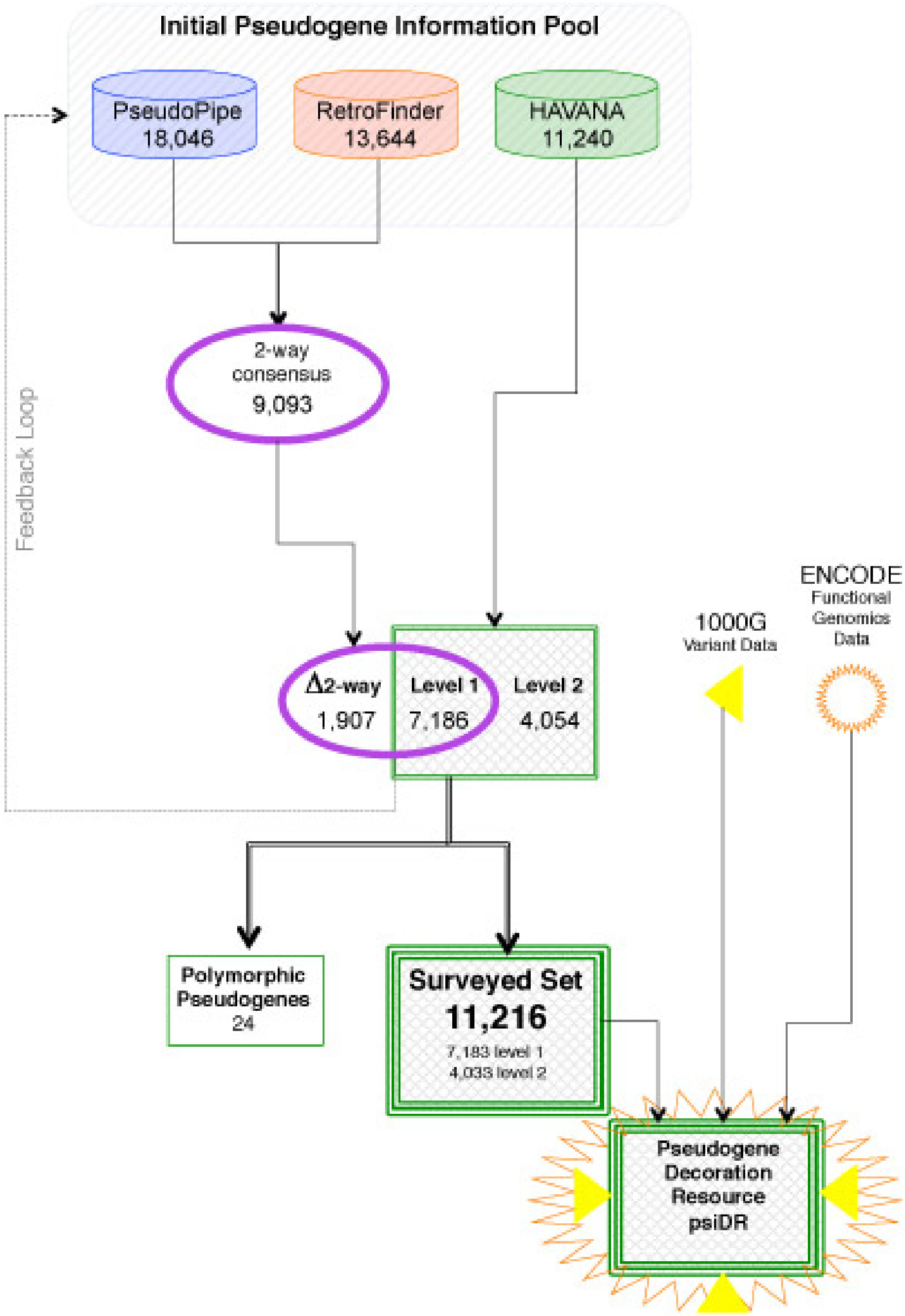


Figure 1. | **Pseudogene annotation flowchart.** A flowchart to describe the GENCODE pseudogene annotation procedure and the incorporation of functional genomics data from the 1000 Genomes Project and ENCODE. This is an integrated procedure including manual annotation done by the HAVANA team and two automated prediction pipelines: PseudoPipe and RetroFinder. The loci that are annotated by both PseudoPipe and RetroFinder are collected in a subset labeled as "2-way consensus", which is further intersected with the manually annotated HAVANA pseudogenes. The intersection results in three subsets of pseudogenes. Level 1 pseudogenes are loci that have been identified by all three methods (PseudoPipe, RetroFinder and HAVANA). Level 2 pseudogenes are loci that have been discovered through manual curation and were not found by either automated pipeline. Delta 2-way contains pseudogenes that have been identified only by computational pipelines and were not validated by manual annotation. As a QC (quality control) exercise to determine completeness of pseudogene annotation in chromosomes that have been manually annotated, 2-way consensus pseudogenes are analysed by the HAVANA team to establish their validity and are included in the manually annotated pseudogene set if appropriate. The final set of pseudogenes are compared with functional genomics data from ENCODE and genomic variation data from the 1000 Genome Project.

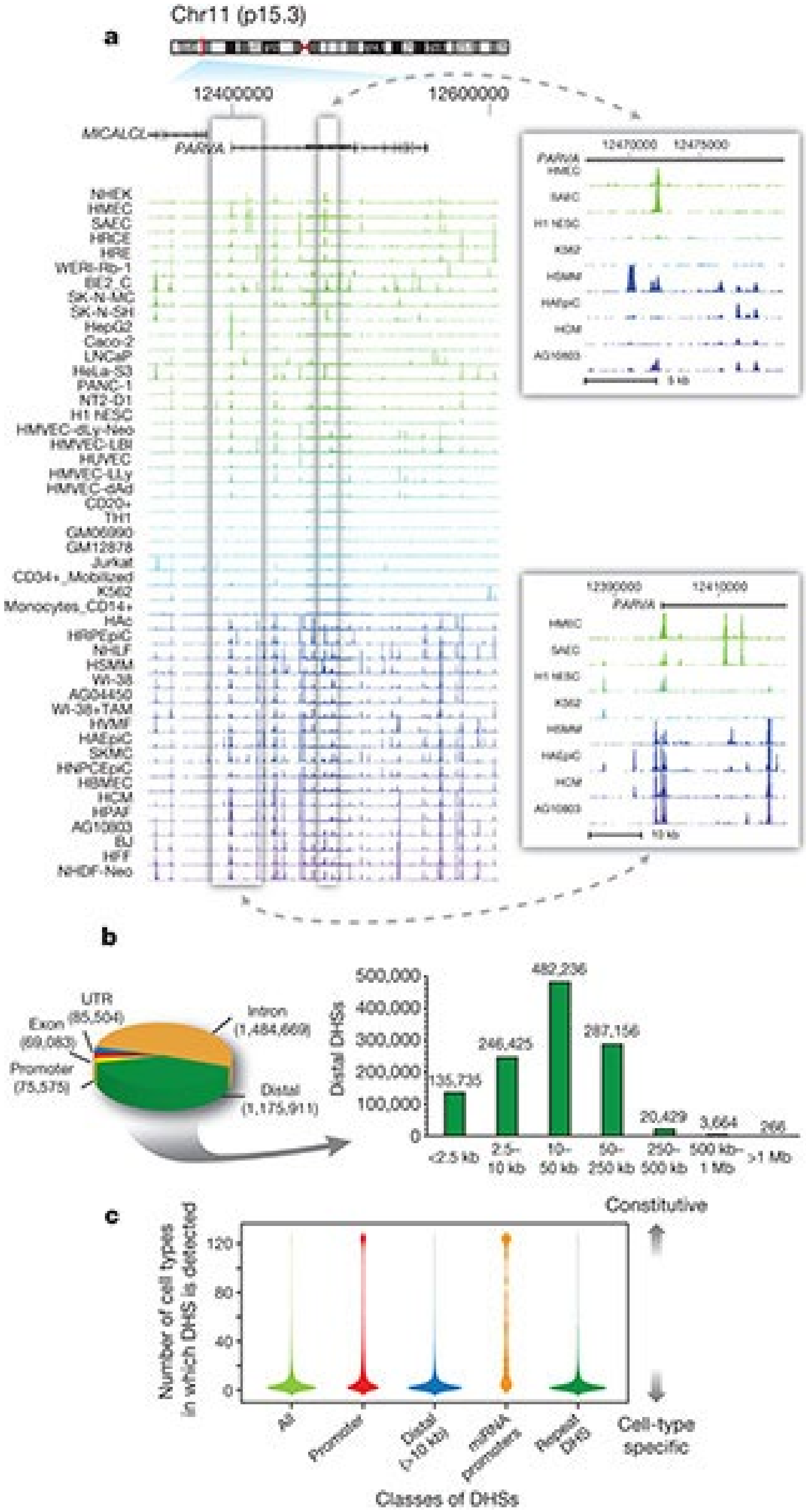


Figure 1 | **General features of the DHS landscape.** (**a**) Density of DNase I cleavage sites for selected cell types, shown for an example ,350-kb region. Two regions are shown to the right in greater detail. (**b**) Left: distribution of 2,890,742 DHSs with respect to GENCODE gene annotations. Promoter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

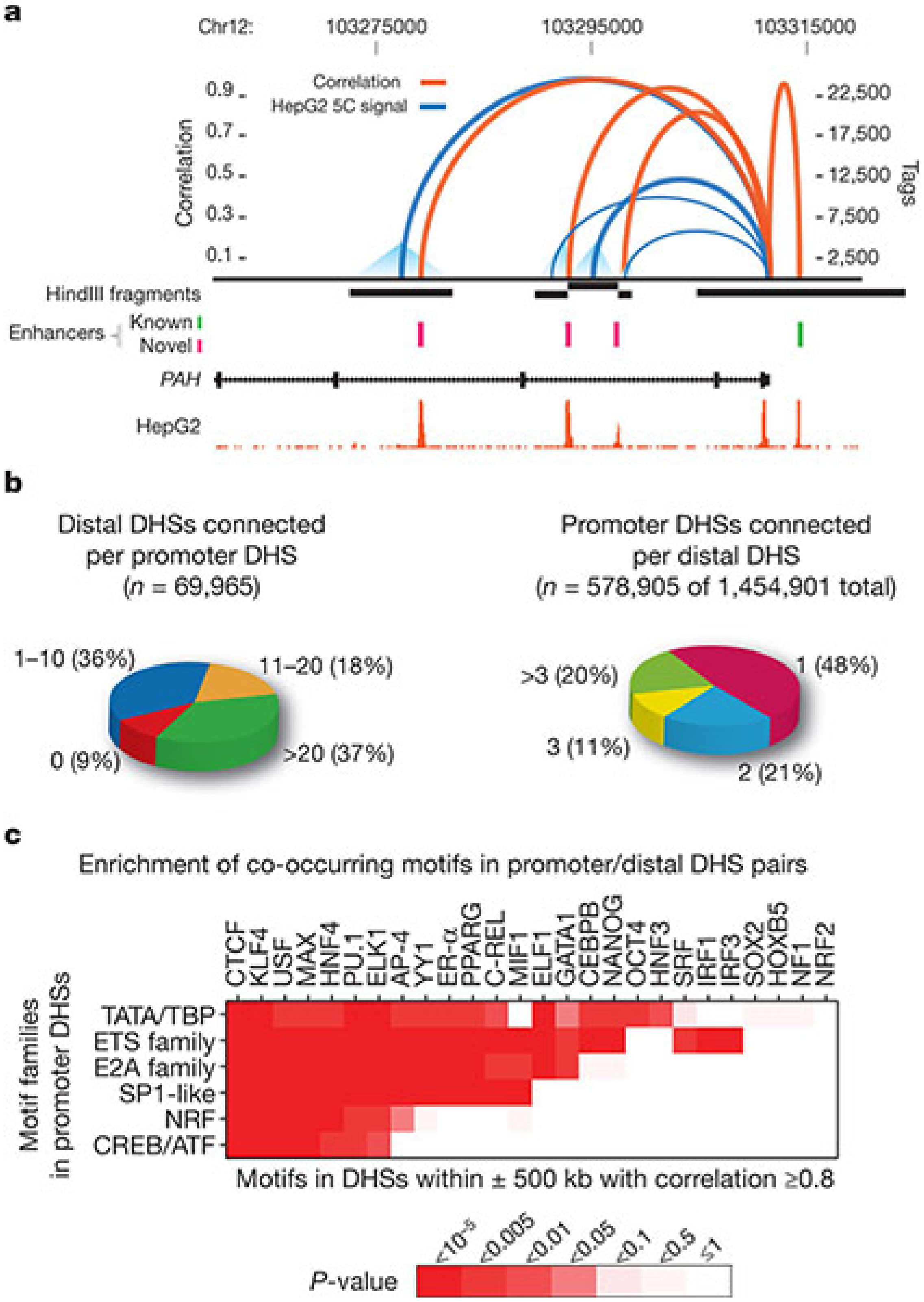


Figure 5 | A genome-wide map of distal DHS-to-promoter connectivity. (a) Cross-cell-type correlation (red arcs, left *y* axis) of distal DHSs and *PAH* promoter closely parallels chromatin interactions measured by 5C-seq (blue arcs, right *y* axis); black bars indicate HindIII fragments used in 5C assays. Known (green) and novel (magenta) enhancers confirmed in transfection assays are shown below. Enhancer at far right is not separable by 5C as it lies within the HindIII fragment containing the promoter. (b) Left: proportions of 69,965 promoters correlated (r>0.7) with 0 to >20 DHSs within 500 kb. Right: proportions of 578,905 non-promoter DHSs (out of 1,454,901) correlated with 1 to >3 promoters within 500 kb. (c) Pairing of canonical promoter motif families with specific motifs in distal DHSs.

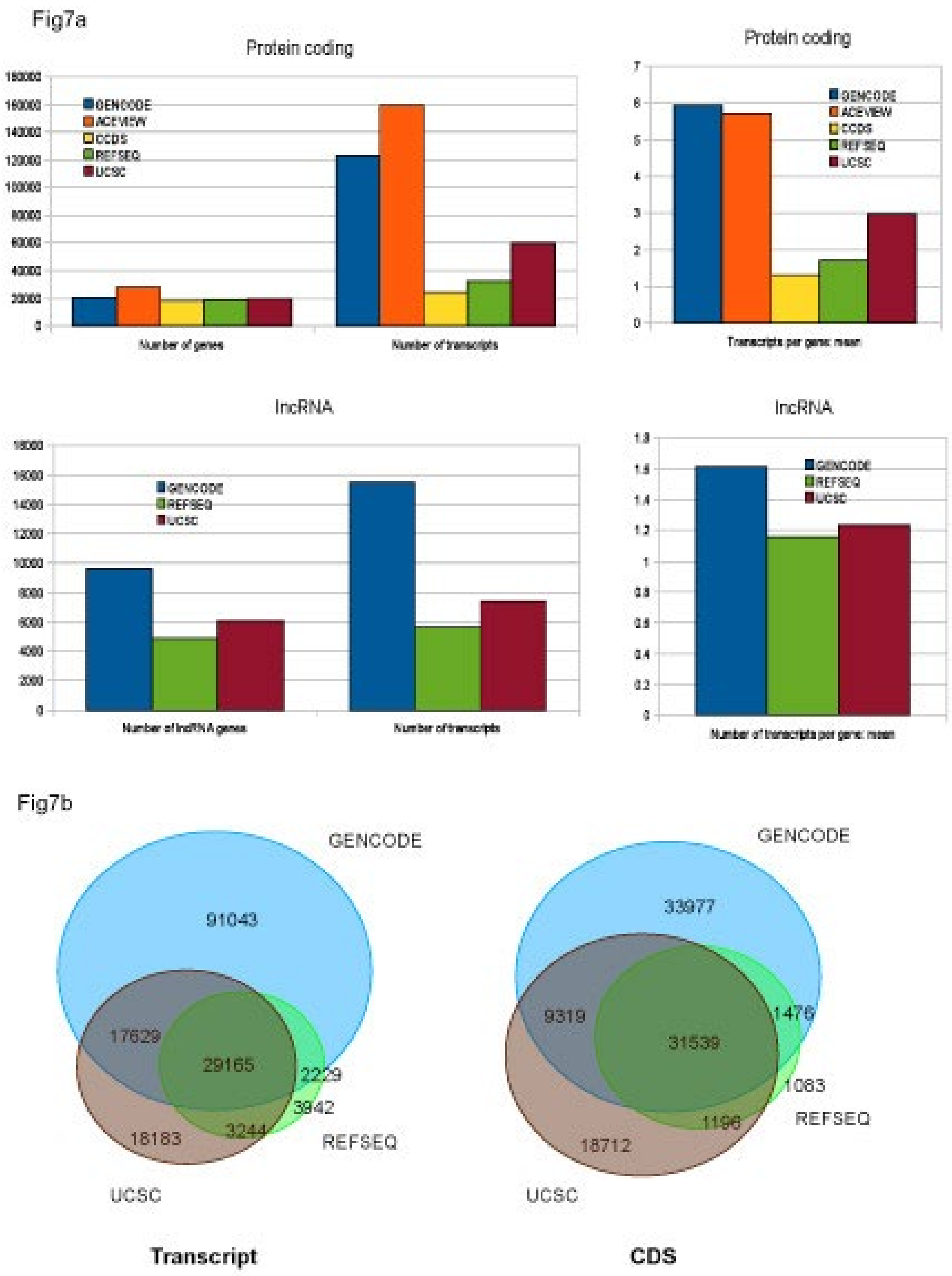


Figure 7 | (a) Comparing different publicly available gene sets. The protein-coding content of five major publicly available gene sets- GENCODE, AceView, consensus coding sequence (CCDS), RefSeq, and UCSC-were compared at the level of total gene number, total transcript number, and mean transcripts per locus. (Blue) GENCODE data; (orange) AceView; (yellow) CCDS; (green) RefSeq; (red) UCSC. The lncRNA content of three of these gene sets-GENCODE, RefSeq, and UCSC-were also compared at the level of total gene number, total transcript number, and mean transcripts per locus. Again, GENCODE data are shown in blue, RefSeq in green, and UCSC in red. (b) Overlap between GENCODE, RefSeq, and UCSC at the transcript and CDS levels. Both protein-coding and lncRNA transcripts of all data sets were compared at the transcript level. Two transcripts were considered to match if all their exon junction coordinates were identical in the case of multi-exonic transcripts, or if their transcript coordinates were the same for mono-exonic transcripts. Similarly, the CDSs of two protein-coding transcripts matched when the CDS boundaries and the encompassed exon junctions were identical. Numbers in the intersections involving GENCODE are specific to this data set, otherwise they correspond to any of the other data sets.

Supplementary Table 7 | Promotor/distal DHS pairs with correlation ≥7Genomic coordinates of all promoter DHSs and distal, non-promoter DHSs within ±500 kb correlated with them at threshold 0.7. Due to the size of this file, we are making it available through the EBI ftp server.This compressed, tab-delimited text file contains 1,672,958 lines of data, for 63,318 distinct promoter DHSs that each have at least one distal DHS connected to it. Each promoter DHS overlaps a TSS, or is the nearest DHS to the TSS in the 5' direction; columns 1-3 contain each promoter DHS's genomic coordinates (hg19). The Gencode gene names are given in column 4. Because distinct gene names can be given to the same TSS, and because distinct TSSs can have the same nearby DHS called as their promoter DHS, data for each promoter DHS is repeated in this file roughly three times on average, with a different gene name for each repetition (there are 207,878 distinct combinations of promoter DHS + gene name in this file). Columns 5-7 contain the genomic coordinates for each distal, non-promoter DHS within 500kb of the promoter DHS given in columns 1-3 that achieves correlation ≤0.7 with it; the correlation between the promoter/distal DHS pair is given in column 8. Distal DHSs appear multiple times in the file when they achieve correlation ≤0.7 with multiple promoter DHSs. Using program sort-bed from the BEDOPS genomic data analysis software suite, from the command line within a Unix system, the set of 578,905 distal DHSs connected with at least one promoter DHS can be extracted into a file named "outfile" by executing the commandcut -f5-7 infile | sort-bed - | uniq > outfilewhere "infile" represents the file genomewideCorrs_above0.7_promoterPlusMinus500kb_withGeneNames_32celltypeCategories.bed8.The first five lines of data are shown below.

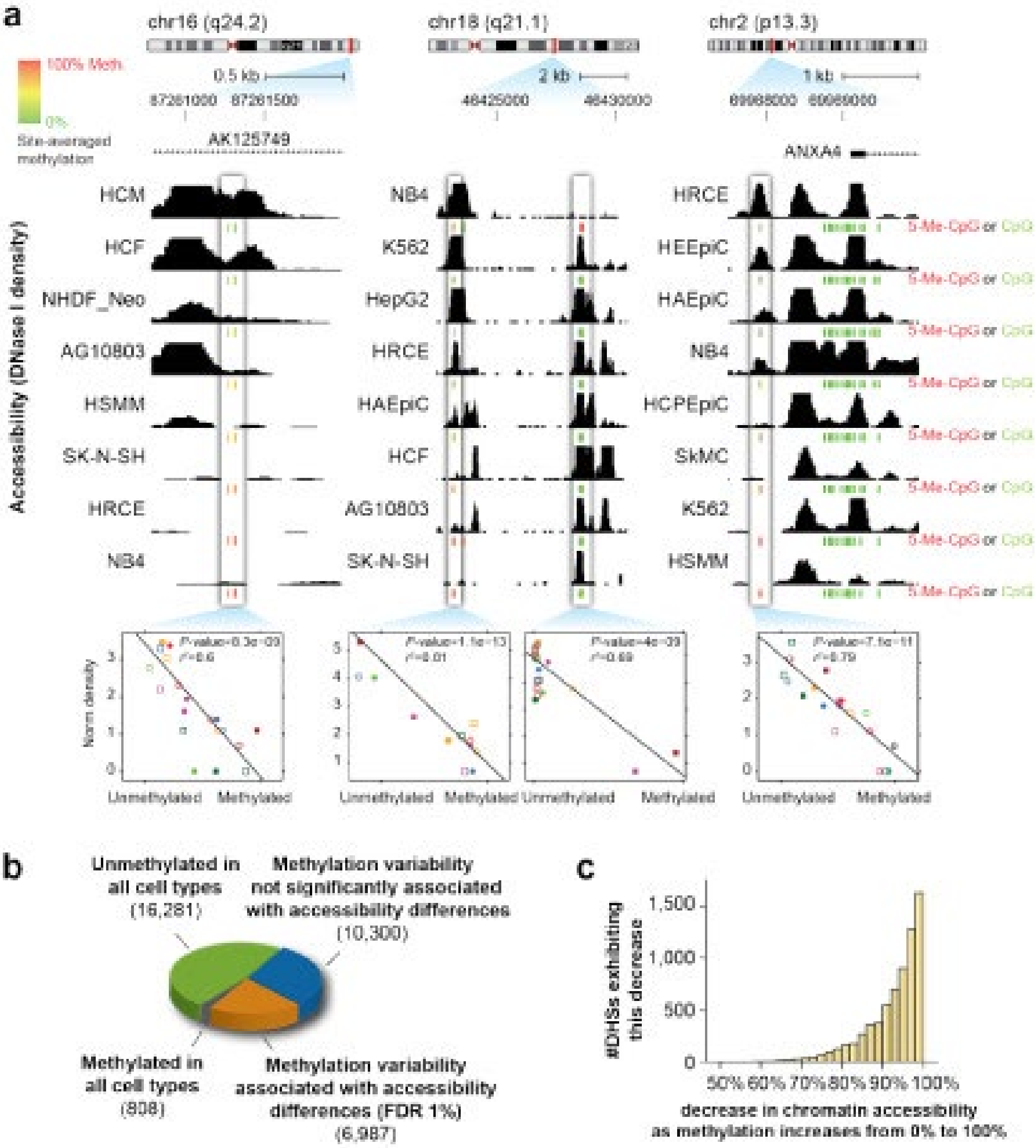
chr1	66660	66810	AL627309.1	chr1	87640	87790	0.87171
chr1	66660	66810	AL627309.1	chr1	118840	118990	0.908176
chr1	66660	66810	AL627309.1	chr1	136960	137110	0.915177
chr1	66660	66810	AL627309.1	chr1	566760	566910	0.731457
chr1	96520	96670	RP11-34P13.8	chr1	237020	237170	0.786171

The workflow used to identify the pseudogenes in this dataset is described in Figure 1. In addition to the 11,216 pseudogenes, the "2-way" consensus set derived from the automated pipeline annotations includes an additional 1,910 pseudogenes (including 3 level 1 polymorphic pseudogenes). As manual annotation is done in a chromosome-by-chromosome fashion, it is not biased relative to any particular genomic feature. Thus, we feel that our "surveyed set" is the best representative of the total pseudogenes complement in the genome.

Transcription initiation requires the binding of multi-protein complexes that position RNA polymerase II[20-23]. Using a modified footprint detection algorithm designed to detect larger features (Supplementary Methods), we scanned the regions upstream from Gencode TSSs and identified highly stereotyped ~80-bp chromatin structure comprising a prominent ~50-bp central DNaseI footprint, flanked symmetrically by ~15-bp regions of uniformly elevated DNaseI cleavage (Fig. 4a). Alignment of per-nucleotide DNaseI cleavage profiles from 5,041 prominent footprints mapped in different K562 promoters highlights the homogeneous, nearly invariant nature of the structure (Fig. 4b).

Plotting evolutionary conservation in parallel with DNaseI cleavage revealed two distinct peaks in evolutionary conservation within the central footprint (Fig. 4c) compatible with binding sites for paired canonical sequence-specific transcription factors. The density of capped analysis of gene expression (CAGE) tags (Fig. 4d; green line) and 5 ends of expressed sequenced tags (ESTs) (Fig. 4d; orange line) relative to the central ~50-bp footprint revealed that, at the vast majority of promoters, RNA transcript initiation localized precisely within the stereotyped footprint. It is notable that the location of this footprint is often offset, typically 5', from many Gencode-annotated TSSs. This probably derives from the incomplete nature of many of the 5' transcript ends used to define TSSs[24].

These data together define a new high-resolution chromatin structural signature of transcription initiation and the interaction of the pre-initiation complex with the core promoter. Indeed, chromatin occupancy of TATA-binding protein (TBP), a critical component of the pre-initiation complex, is maximal precisely over the centre of the 50-bp footprint region (Supplementary Fig. 9a). Sequence analysis of the two conservation peaks within the 50-bp footprint identified motifs for GC-box-binding proteins such as SP1 and, less frequently, other general transcription factors (though with the notable absence of TATA motifs) (Supplementary Fig. 9b), indicating that TBP (and potentially other pre-initiation complex components) interacts preferentially with general transcriptional factors bound to GC-box-like features in the central footprinted region. The results are therefore consistent with a model in which a limited number of sequence-specific factors function both to prime the chromatin template for recruitment of RNA polymerase II and to guide transcriptional positioning.



Supplementary Figure 14 | Interaction and GO class enrichments via signal-vector correlation. (a) Further examples of association between methylation and accessibility. Data tracks show DNase I sensitivity in selected cell types. Green bars, CpG is 0% methylated; yellow, 50% methylated; red, 100% methylated. Association is quantified in the plots below the tracks. Each point in the graph represents one of 19 cell-types (a susbset of which is represented in the tracks). *X*-axis is the percent methylation of the site in that cell-type; *y*-axis is the normalised DNaseI tag density at the site in that cell type. In each example, accessibility (*y*-axis) quantitatively decreases as methylation increases (left to right). (b) Global characterisation of the effect of methylation on chromatin accessibility, surveyed at 34,376 DHSs with RRBS data. 40% of sites with variable methylation across cell-types were associated with differences in chromatin accessibility. (c) In cell lines with methylated DHSs, site accessibility was reduced on average by 95%. Shown are sites where increased methylation was significantly associated with decreased accessibility (= 97% of all sites in the orange slice shown in (b)).

RT-PCR-seq to substantiate RNA-seq predictions

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

RT-PCR-seq to identify novel transcript isoforms

To investigate how many unsubstantiated exon-exon junction exhibit non targeted splice sites in the amplified amplimers, we remapped all unmappable reads within the primer boundaries with GEM (http://sourceforge.net/apps/mediawiki/gemlibrary/index.php?title=The_GEM_library), a sequence aligner that allows, contrary to Tophat (Trapnell *et al.* 2009) split-mapping without prior prediction of genomic islands corresponding to exons. This is crucial as in RT-PCR-seq one of the primers is intentionally designed very close to the assessed junction (Figure 1B, Methods). To quantify the fraction of PCR amplifications that allowed identification of novel isoforms we implemented a specific scoring method (see Methods). We found 1119 (11%, n=10,162) unannotated transcript models including 644 new internal exons and 568 novel splicing events within known exons (new exons that intersect not more than 80% of the length of a GENCODE annotated exon, see Methods; note that some assessed genomic intervals have both class of novel exons). The vast majority of these novel exons (86%, 1046/1212) present canonical donor and acceptor sites. 44 and 83% of the internal new exons are overlapped over 90% of their length by sequencing reads of the deep transcriptome profiling of 16 different human tissues (4.99 billion individual sequence reads from adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testis, thyroid and white blood cells polyA+ RNA) and 15 human cell lines (5.55 billion sequences from A549, AG04450, BJ, GM12878, H1-hESC, HMEC, HSMM, HUVEC, HeLa-S3, HepG2, K562, MCF7, NHLF, NHEK and SK-N-SH polyA+ RNA) generated by the Illumina "Human Body Map" (HBM) project and ENCODE (The ENCODE Consortium 2012; Djebali *et al.* 2012b) confirming that they are new transcript models. Likewise, 45 novel internal exons within protein coding loci are supported by newly released ESTs and/or cDNAs. Their splice sites (44 canonical, 1 non-canonical) are conserved across several species ((Nitsche *et al.* 2012), see also http://splicemap.bioinf.uni-leipzig.de/). 34 of these exons are incorporated in protein coding transcripts, the remaining 11 are integrated in processed transcripts. Some examples are presented in Figure 3-4 and Supplementary Figure S1C-S1D. A large fraction of these new exons is tissue specific (69%), but some are detected ubiquitously (Figure 3B-4B) (e.g. 6% were identified in at least 5 tissues).

All internal canonically-spliced new exons of protein coding genes (n=313) were subsequently manually annotated by the GENCODE pipeline. They are generally poorly conserved across vertebrates as shown by the distribution of their phastCons scores (Figure 5A). They do not overlap repeats more than intergenic and intronic sequences as repeatedly shown for lineage specific exons (reviewed in (Keren *et al.* 2010)) (Supplementary Figure S2). They can be subdivided into 70 new coding (22%), 173 new nonsense-mediated decay (NMD, 55%), 55 novel UTR (18%) and 15 new non-coding exons (5%) (Supplementary Table S4). Whereas it is difficult to draw general conclusion about the functionality of these new exons, some interesting cases could be pinpointed. For example, a new exon in the *BAD* gene interrupts one pro-apoptotic Pfam domain (Bcl-2_BAD, PF10514) but inserts another domain (GVQW, PF13900) commonly found in caspases, a family of proteins crucial to the apoptotic pathway (Figure 6A). The two new NMD-inducing "poison" (Lareau *et al.* 2007) prone exons found in the *NR1H4* locus are highly-specific to liver possibly controlling expression of that gene in this tissue (Figure 6B). Likewise, we identified two new mutually exclusive 5'UTR *ECI2* exons (Figure 6C). Some of the novel exons, especially within the coding, NMD and UTR categories are evolutionary-conserved (see outliers in Figure 5B). For example, we identified a new highly conserved exon within the *KIAA0528* gene (Figure 6D). Its acceptor and donor sites are conserved back to medaka, while the encoded peptide is highly conserved back to the anolis lizard (Supplementary Figure S3). We conclude that RT-PCR-seq can be used to further improve the current annotation and discover new gene structures.

To fully exploit this information, the scientific community requires a reliable coding and non-coding gene catalogue, compilation of which was assigned, within ENCODE, to the GENCODE consortium (Harrow *et al.* 2012). With more than 51,096 genes (20,026 coding and 31,070 non-coding; GENCODE version 8, March 2011), this manually curated annotation is richer than any previously available annotation (e.g. UCSC genome browser (Fujita *et al.* 2011)). We demonstrate here that it is of extremely high quality, as even its lower confidence gene and transcript models can be experimentally validated using RT-PCR-seq (79.3% and 80.7% validation rate with more or less conservative criteria [...]). It does not, however, fully represent the complexity of the human transcriptome, since we identify novel internal exons in more than 11% of the genomic intervals we interrogated with RT-PCR-seq. As GENCODE coding and long non-coding transcripts have an average of 4.3 and 2.2 exons, respectively (Harrow *et al.* 2012), we can conservatively estimate that about 18% of the annotated genome loci have yet unrecognized exons.

Approximately 3% ($n = 75,575$) of DHSs localize to transcriptional start sites (TSSs) defined by Gencode10 and 5% ($n = 135,735$, including the aforementioned) lie within 2.5 kilobases (kb) of a TSS. The remaining 95% of DHSs are positioned more distally, and are roughly evenly divided between intronic and intergenic regions (Fig. 1b). Promoters typically exhibit high accessibility across cell types, with the average promoter DHS detected in 29 cell types (Fig. 1c, second column). By contrast, distal DHSs are largely cell selective (Fig. 1c, third column).

From examination of DNaseI profiles across many cell types we observed that many known cell-selective enhancers become DHSs synchronously with the appearance of hypersensitivity at the promoter of their target gene (Supplementary Fig. 13). To generalize this, we analysed the patterning of 1,454,901 distal DHSs (DHSs separated from a TSS by at least one other DHS) across 79 diverse cell types (Supplementary Methods and Supplementary Table 6), and correlated the cross-cell-type DNaseI signal at each DHS position with that at all promoters within ±500 kb (Supplementary Fig. 14a). We identified a total of 578,905 DHSs that were highly correlated ($r > 0.7$) with at least one promoter ($P < 10\text{-}100$), providing an extensive map of candidate enhancers controlling specific genes (Supplementary Methods and Supplementary Table 7). To validate the distal DHS/enhancer-promoter connections, we profiled chromatin interactions using the chromosome conformation capture carbon copy (5C) technique[31]. For example, the phenylalanine hydroxylase (*PAH*) gene is expressed in hepatic cells, and an enhancer has been defined upstream of its TSS (Fig. 5a). The correlation values for three DHSs within the gene body closely parallel the frequency of long-range chromatin interactions measured by 5C. The three interacting intronic DHSs cloned downstream of a reporter gene driven by the *PAH* promoter all showed increased expression ranging from three- to tenfold over a promoter-only control, confirming enhancer function.

We next examined comprehensive promoter-versus-all 5C experiments performed over 1% of the human genome[32] in K562 cells. DHS-promoter pairings were markedly enriched in the specific cognate chromatin interaction ($P < 10\text{-}13$, Supplementary Fig. 14b). We also examined K562 promoter-DHS interactions detected by polymerase II chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)24, which quantifies interactions between promoter-bound polymerase and distal sites. The ChIA-PET interactions were also markedly enriched for DHS-promoter pairings ($P < 10\text{-}15$, Supplementary Fig. 14c). Together, the large-scale interaction analyses affirm the fidelity of DHS-promoter pairings based on correlated DNaseI sensitivity signals at distal and promoter DHSs.

Most promoters were assigned to more than one distal DHS, indicating the existence of combinatorial distal regulatory inputs for most genes (Fig. 5b and Supplementary Table 7). A similar result is forthcoming from large-scale 5C interaction data[32]. Surprisingly, roughly half of the promoter-paired distal DHSs were assigned to more than one promoter (Fig. 5b and Supplementary Methods), indicating that human *cis*-regulatory circuitry is significantly more complicated than previously anticipated, and may serve to reinforce the robustness of cellular transcriptional programs.

The number of distal DHSs connected with a particular promoter provides, for the first time, a quantitative measure of the overall regulatory complexity of that gene. We asked whether there are any systematic functional features of genes with highly complex regulation. We ranked all human genes by the number of distal DHSs paired with the promoter of each gene, then performed a Gene Ontology analysis on the rank-ordered list. We found that the most complexly regulated human genes were markedly enriched in immune system functions (Supplementary Fig. 14d), indicating that the complexity of cellular and environmental signals processed by the immune system is directly encoded in the *cis*-regulatory architecture of its constituent genes.

We compared the composition of annotation across the five major gene sets publicly available in UCSC, GENCODE, CCDS, RefSeq, UCSC and AceView. Both the number of protein-coding loci and transcripts at those loci were investigated. The CCDS set has the lowest number of protein-coding loci and alternatively spliced transcripts since it is a high quality conservative gene set derived from RefSeq and Ensembl/HAVANA gene merge (Pruitt *et al.* 2009). In CCDS every splice site of every transcript must agree in both the RefSeq and Ensembl/Havana gene set and all transcripts must be full-length. While the number of protein-coding loci in RefSeq, GENCODE and UCSC is comparable, AceView has approximately 20,000 more coding loci. One likely source of inflation is the predisposition for AceView to add a CDS to transcript model, and hence create novel loci from lncRNAs and pseudogenes eg PTENP. AceView predicts 31,057 single exon loci compared with 1,724 in GENCODE, 3,234 RefSeq and 4,731 in UCSC genes. Excluding single exon loci predicted by AceView from this analysis, the number of AceView gene loci is much closer to the number in other gene sets (see Figure 7a).

Using next generation sequencing to find novel protein-coding and lncRNA genes outside GENCODE

To identify novel coding and non-coding genes represented in RNA-seq data, we studied transcript models reconstructed using Exonerate (Howald *et al.*, 2012) and Scripture (Guttman *et al.* 2010), based on the high depth HBM transcriptomic data from 16 tissues made publicly available from Illumina (ArrayExpress accession: E-MTAB-513; ENA archive: ERP000546).

Assessing coding potential of RNA-seq models using PhyloCSF

We analysed the resulting transcripts that did not overlap any GENCODE loci for coding potential using PhyloCSF (Lin *et al.* 2011), which examines evolutionary signatures within UCSC vertebrate alignments including 33 placental mammals. There were 136 Ensembl HBM models with positive PhyloCSF scores out of a total of 3,689 loci, although only five of these had sufficient support for manual reannotation as coding genes (see Supplementary Table 8). The remaining 131 transcripts showed varying quality and evidence; around 50% overlap novel processed transcripts and could be a result of misalignment of reads or actual expressed pseudogenes. Two hundred Scripture transcript predictions that were outside GENCODE but had high PhyloCSF scores were also manually examined. Of these 15 were added as novel loci and only 9 were annotated as coding genes (see Supplementary Table 9) and will be added to the next release of GENCODE). Considering the depth of reads of the HBM data (averaging over a billion read depth) from the 16 different tissues, we have not identified many missing coding genes based on PhyloCSF. Indeed, since 3,127 HBM Ensembl genes consist of only two exons it is highly likely these constitute new lncRNAs we have not yet annotated and will be merged into a later release of GENCODE.