SUPPLEMENTARY INFORMATION

doi:10.1038/nature21374

An atlas of human long non-coding RNAs with accurate 5' ends

by Division of Genome Technologies, RIKEN, Yokohama, Japan (2017)

TABLE OF CONTENTS

A | Supplementary Notes 2
1 | Rationale and Implementation of TIEScore 2
2 | Benchmarking the Performance of TIEScore 4
3 | TIEScore Cutoffs and TSS Validation Rates 6
4 | Definition of Genes, Coding Potential and Comparison with GENCODEv25 7
5 | Directionality, Exosome Sensitivity and Transcript Properties 9
6 | Sequence Features at LncRNA TSS Supported by Different DHS Types 11

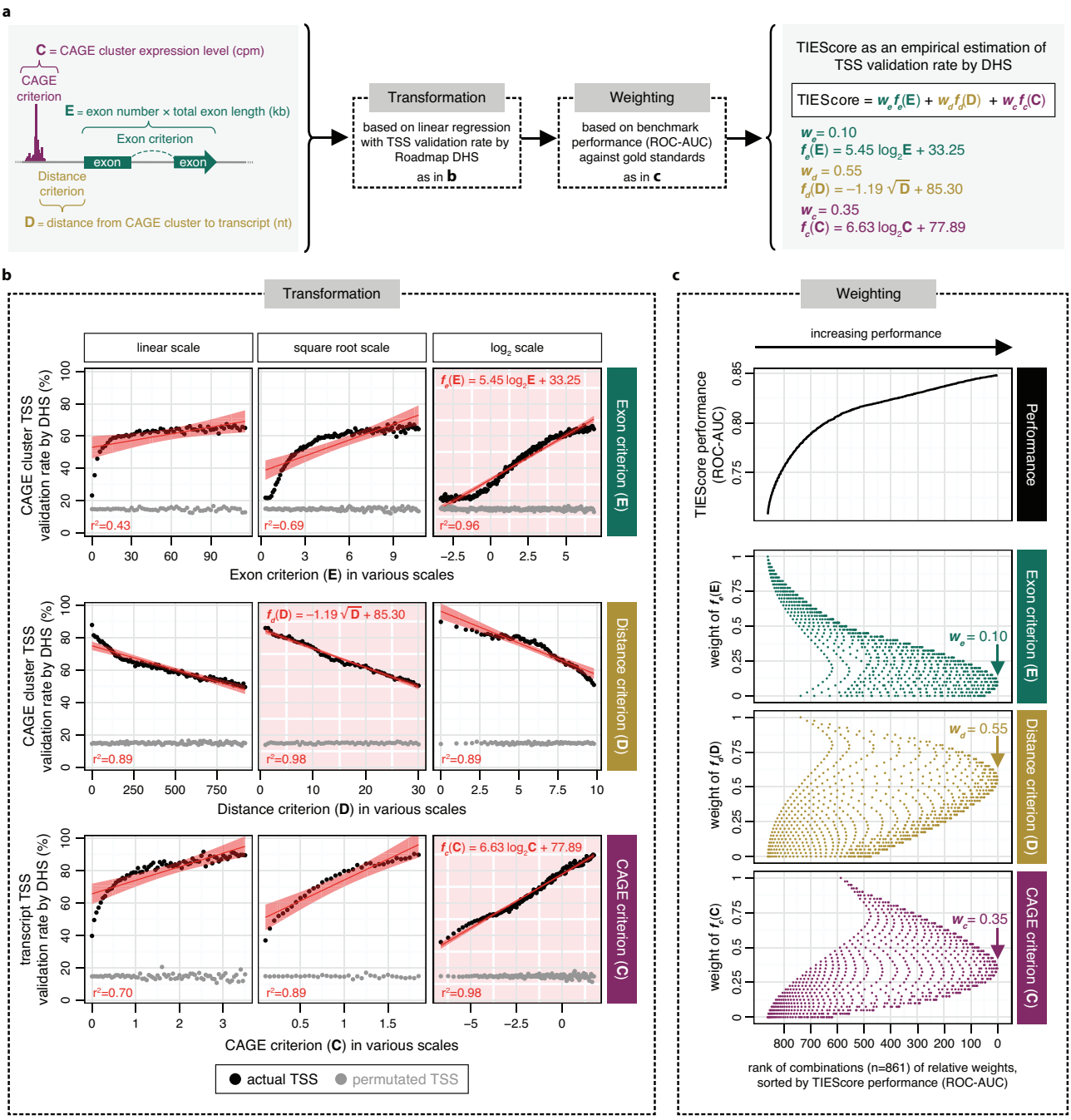
B | Online Resources 12
1 | Assembly, Expression Atlas and other Resources 12
2 | Overview of FANTOM CAT Browser 13
3 | Use Case 1: Download a List of Novel e-lncRNAs 14
4 | Use Case 2: Explore LncRNAs with Conserved Exons and Implicated in eQTL 15
5 | Use Case 3: Explore LncRNAs Enriched in Classical Monocytes 17
6 | Use Case 4: Explore Cell Types Associated with Crohn's Diseases 19

C | Supplementary Tables 20
1 | CAGE Library Information 20
2 | RNA-seq Library Information 20
3 | FANTOM CAT Gene Information 20
4 | Directionality, Exosome Sensitivity and Transcript Properties 20
5 | FANTOM CAT Genes in LncRNAdb 20
6 | Conservation of TIR and Exon 20
7 | Transposons at TIR 20
8 | Orthologous Transcription 20
9 | Expression Levels and Specificity in Primary Cell Facets 20
10 | Sample Ontology Information 20
11 | Gene Association with Cell Types 20
12 | Trait Information 20
13 | Gene Association with Traits 20
14 | Curation of Cell type and Trait Pairs 21
15 | Genes Involved in Cell Type and Trait pairs 21
16 | eQTL-linked lncRNA and mRNA Pairs 21
17 | Gene-based Functional Evidence 21
18 | Grouping of Samples for Differential Expression 21
19 | Differential Expression Results 21

D | References 22

A | Supplementary Notes

1 | Rationale and Implementation of TIEScore



Supplementary Fig. 1 | Rationale and Implementation of TIEScore. **a**, For each pair of transcript and CAGE cluster, TIEScore is calculated as the sum of values from exon (**E**), distance (**D**) and CAGE (**C**) criteria, which were transformed by f_e(**E**), f_d(**D**) and f_c(**C**) as in **b** and weighted by w_e, w_d and w_c as in **c**, respectively. The resulting TIEScore can be interpreted as an empirical estimation of TSS validation rate by DHS. **b**, Transformation of TIEScore criteria values. We investigated the correlation between TIEScore criteria values and TSS validation rate by DNaseI hypersensitivity sites (DHS). The criteria values were transformed into various scales (**columns**) and plotted against TSS validation rate by DHS. A TSS is 'validated' if it overlaps a DHS. Within each bin of values at **X-axis**, the percentage of validated TSS was calculated (i.e. TSS validation rate). Same number of positions was randomly sampled from the unannotated genomic regions (permutated TSS). Linear regression (**red line**, 99.99% confidence intervals) was performed on the actual TSS and r^2 was indicated. The plot with highest r^2 within each TIEScore criteria was highlighted in **pink** and the corresponding linear regression functions (i.e. f_e (**E**), f_d (**D**) and f_c (**C**)) were used to transform the TIEScore criteria values. **c**, Weighting of TIEScore criteria values. TIEScore is defined as the weighted sum of the transformed TIEScore criteria values.

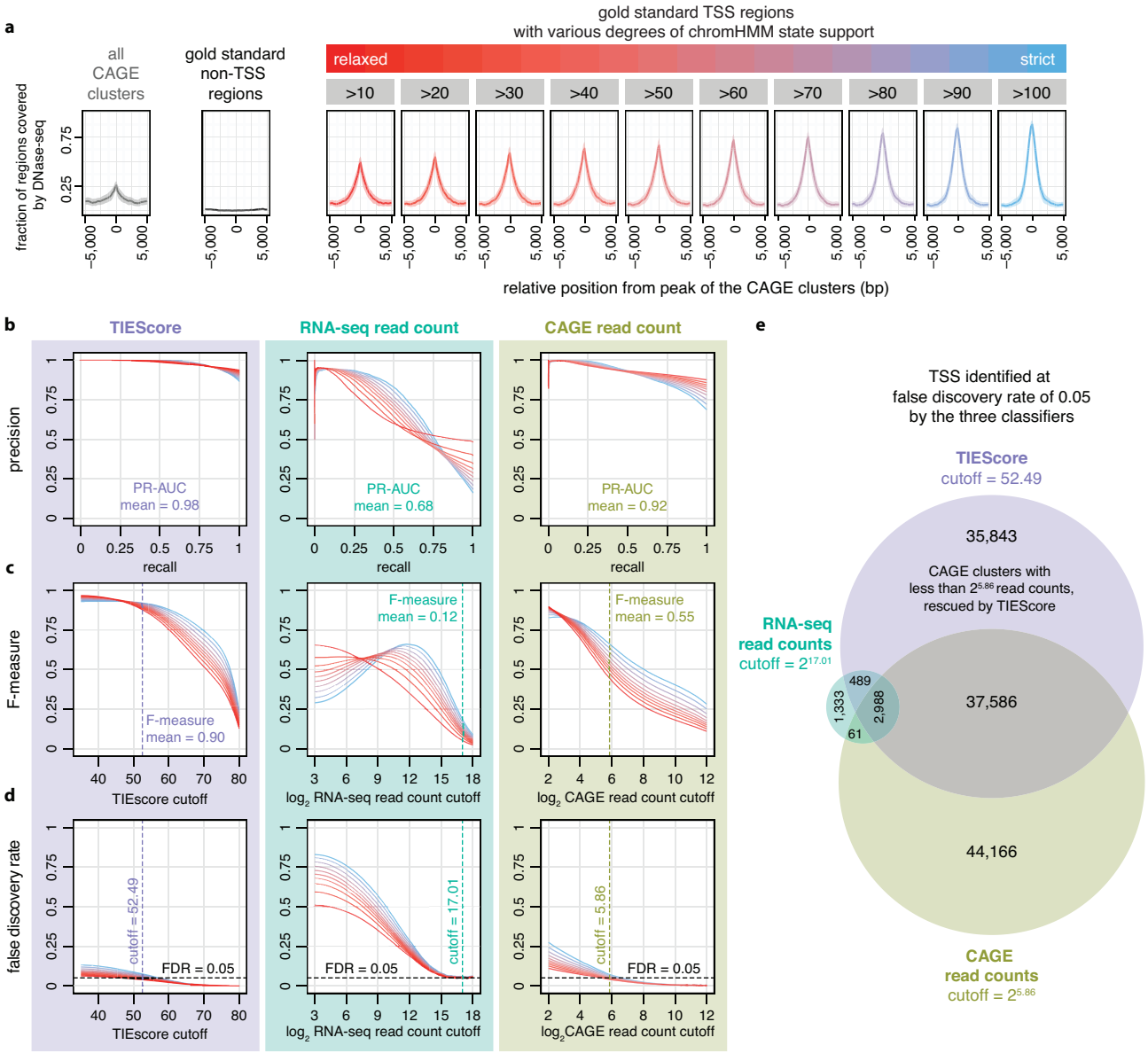
1.1 TIEScore Rationale: Transformation and Weighting of Criteria Values

Transcription Initiation Evidence Score (TIEScore) is a custom metric that evaluates the properties of a pair of CAGE cluster and transcript model to quantify the likelihood that the corresponding CAGE transcription start site (TSS) being genuine (**Supplementary Fig. 1a**). We reasoned lowly abundant transcripts are likely to be insufficiently covered by RNA-seq reads and thus their transcript models are more likely to be incomplete and truncated into shorter and less spliced models[1]. This is supported by the observation that the product of exon number and total exon length (kb) of transcript models (i.e. Exon criterion, **E**) are highly correlated with the percentage of TSS validated by DHS (i.e. DHS validation rate, **Supplementary Fig. 1b**). Also, lowly abundant CAGE clusters that are distant from transcript 5' ends may represent degradation products of longer transcripts, previously referred to as 'exon painting'[2]. This is supported by the observation that the expression levels of CAGE clusters and their distances to the closest transcript model (i.e. CAGE criterion, **C** and Distance criterion, **D**) are correlated with DHS validation rate (**Supplementary Fig. 1b-c**). We therefore devised TIEScore (**Supplementary Fig. 1a**), a metric integrating these criteria to address inaccurately identified 5' ends of transcripts and build transcript models with well-supported transcription initiation evidence. First, we explored the linearity of DHS validation rate across these criteria values using various scales based on linear regression (**Supplementary Fig. 1b**). For each criterion, the scale that yields the highest r^2 was chosen (**Supplementary Fig. 1b**) and its value was then transformed into DHS validation rate based on the corresponding linear regression functions (i.e. $f_e(\mathbf{E})$, $f_d(\mathbf{D})$ and $f_c(\mathbf{C})$) as in **Supplementary Fig. 1b**. TIEScore is defined as the weighted sum of these transformed values. To optimize the weights of these criteria, we evaluated the performance of TIEScore for discrete combinations of weights (n=861 combinations, i.e. three weights at grid of 0.025 with sums equal to 1, **Supplementary Fig. 1c**). The performance of TIEScore was measured in terms of the area under receiver operating characteristic (ROC) curve (ROC-AUC)[3] in 70 matched CAGE and RNA-seq libraries (see details in **Supplementary Note 2**) and the combination of weights which yielded the highest AUC-ROC was chosen (i.e. w_e, w_d and w_c). TIEScore is thus calculated as: $w_e f_e(\mathbf{E}) + w_d f_d(\mathbf{D}) + w_c f_c(\mathbf{C})$, which can be interpreted as an empirical estimation of TSS validation rate by DHS.

1.2 Implementation of TIEScore in Meta-assembly of FANTOM CAT

TIEScore was evaluated for each pair of CAGE clusters and transcript models within 1kb. Each transcript model was assigned to the CAGE cluster yielding the highest TIEScore and its 5' end was adjusted to the most prominent TSS of the CAGE cluster. Transcripts with no CAGE peaks within 1kb, and CAGE peaks without transcripts within 1kb, were therefore discarded. This associated each retained transcript with a CAGE cluster, and each retained CAGE cluster with one or more transcripts. A CAGE cluster is then assigned the highest TIEScore yielded from its associated transcripts. TIEScore was first applied to each of the five transcript model collections separately (with 1,897 CAGE libraries, **Supplementary Table 1**) and then merged into a non-redundant transcript set (referred to as raw FANTOM CAGE associated transcriptome (CAT)). Specifically, the transcript models from GENCODEv19 were used as the initial reference to sequentially overlay onto them the transcripts from the other four collections, in sequence of Human BodyMap 2.0[4], miTranscriptome[5], ENCODE[6] and FANTOM5 RNA-seq assembly (this study). In each overlay, the query transcript models that share 1) 5' ends within ±50bp, 2) exact same splicing junctions, 3) 3' ends within ±2,000bp, with the reference transcript models were discarded as redundant. The non-redundant query transcript models were then added as the new reference set for the next round of overlay. For each CAGE cluster, the maximum TIEScore after all four rounds of overlays was taken as its TIEScore. The TIEScore of raw FANTOM CAT was benchmarked against gold standards (with N=50, see details in **Supplementary Note 2**). The TIEScore cutoffs were chosen as in **Supplementary Note 3**.

2 | Benchmarking the Performance of TIEScore



Supplementary Fig. 2 | Benchmarking the Performance of TIEScore. Seventy samples with matched CAGE and RNA-seq libraries were used for TIEScore benchmarking (**Supplementary table 1**). **a,** DHS coverage of gold standard TSS regions. All CAGE clusters (**1st column**) refer to all CAGE clusters in raw FANTOM CAT. Gold standard non-TSS regions (**2nd column**) and TSS regions (**3rd column**) were defined using chromatin states among 127 epigenome datasets from the Roadmap Epigenomics Consortium[7] and CAGE clusters identified in FANTOM5[8–10]. Gold standard TSS regions were defined at various degrees of chromatin state support (from relaxed to strict, **3rd to last column**). **Y-axis**: fraction of the corresponding regions covered by DNase-seq peaks. **Line and shaded area**: signal summarized per window of 200bp, **solid line** and **shaded area** represent the median and quartiles of 127 epigenome datasets at the corresponding window. In **b, c** and **d,** benchmarking of the performance of TIEScore, CAGE read count and RNA-seq read count was repeated with 10 sets of gold standard TSS regions defined at various levels of stringency as in **a** with same color scale. **b,** Precision and recall curve (PR curve). The area under PR curve (PR-AUC)[3] is used as a measure of classifier performance in identifying genuine TSS. PR-AUC[3] of TIEScore is significantly higher than that of the other 2 classifiers across various stringencies of gold standard TSS definition ($P<0.05$, paired Student's *t*-test). In **c** and **d, vertical dashed lines**: classifier cutoffs with the false discovery rate (FDR) of 0.05 at gold standard TSS>50 chromatin state support. **c**, F-measure versus classifier cutoffs. F-measure[3] was calculated as the harmonic mean of precision and recall and is used as a measure of classifier accuracy in identifying genuine TSS. At FDR of 0.05 (**vertical dashed lines**), TIEScore outperforms (higher F-measure values) the other two classifiers. **d,** FDR versus classifier cutoffs. Intersections of the curve and the **horizontal dashed lines** refer to the classifier cutoffs for achieving FDR of 0.05 (**vertical dashed lines**). **e**, Overlap of TSS identified by the three classifiers at cutoffs for achieving FDR of 0.05.

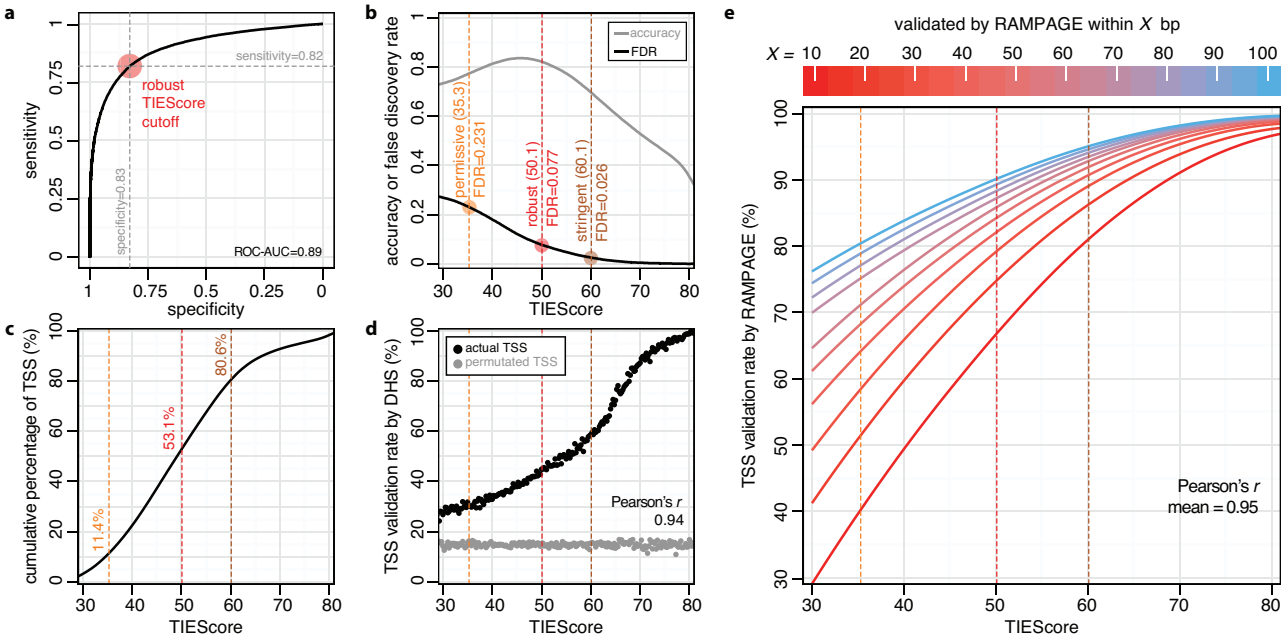
2.1 Definition of gold standard TSS and non-TSS regions

To assess how well TIEScore identifies genuine TSS compared to using CAGE or RNA-seq derived information alone, we first defined sets of gold standard TSS and non-TSS regions using epigenome datasets from the Roadmap Epigenomics Consortium[7] (http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state). Firstly, chromatin states[11] TssA (Active TSS), Enh (Enhancers), TssBiv (Bivalent/Poised TSS) and EnhBiv (Bivalent Enhancer) were referred to as 'pro-TSS-states'; and chromatin states[11] Tx (Strong transcription), Het (Heterochromatin) and Quies (Quiescent/Low) were referred to as 'non-TSS-states'. A CAGE cluster was defined as a gold standard TSS region when its prominent TSS overlaps a pro-TSS-state in N of the 127 epigenomic datasets[7] and overlaps no non-TSS-states, where N reflects the stringency of gold standard definition. Conversely, a CAGE cluster was defined as a gold standard non-TSS region when its prominent TSS overlaps a non-TSS state in at least 120 of the 127 epigenome datasets[7] and overlaps no pro-TSS-states. These gold standard regions were then used to assess TIEScore performance in distinguishing TSS from non-TSS regions (**Supplementary Fig. 2a**).

2.2 Performance of TIEScore versus CAGE or RNA-seq data alone

Using 70 samples with matched CAGE and RNA-seq libraries (**Supplementary Table 2**), the performance of TIEScore in identifying genuine TSS was compared against CAGE or RNA-seq data alone, in identification of genuine TSS (**Supplementary Fig. 2**). First, TIEScore was applied to a subset of CAGE clusters with at least 3 reads (sum among the 70 samples) and the FANTOM5 RNA-seq assembly. A set of universal TSS regions (n=1,011,254) was then generated across the three sets of TSS (i.e. combined, and CAGE or RNA-seq alone) by merging the transcripts 5' ends in the FANTOM5 RNA-seq assembly within ±25nt into RNA-seq TSS regions and then by merging these RNA-seq TSS regions with the CAGE clusters within ±100nt. Gold standard TSS and non-TSS regions for these universal TSS regions were defined using the midpoint of these universal TSS regions instead of prominent TSS in CAGE clusters. Ten sets of gold standard TSS regions were defined at various stringencies as described above, with N=10 to 100, at step of 10. For each of the universal TSS regions, its support (i.e. score) from each of the three datasets was then calculated as, 1) combined: maximum TIEScore of all associated CAGE clusters; 2) CAGE only: maximum number of CAGE reads of all associated CAGE clusters; 3) RNA-seq only: maximum number of RNA-seq reads of all associated RNA-seq TSS regions. (Note: each RNA-seq TSS region is represented by the sum of RNA-seq reads of its associated transcripts as estimated in Sailfish[12]). The score performance for each of the three TSS datasets was assessed using R packages ROCR[13] and PRROC[3]. Specifically, the sensitivity, specificity, precision, recall, FDR and F-measure at various score cutoffs were calculated using ROCR[13], and the ROC-AUC and PR-AUC were calculated using PRROC[3]. In terms of precision, recall and F-measure (**Supplementary Fig. 2b-c**), TIEScore substantially outperformed both CAGE only and RNA-seq only based approaches. In **Supplementary Fig. 2d**, cutoffs values of various classifiers at FDR of 0.05 were indicated. In **Supplementary Fig. 2e**, RNA-seq read count identified appreciably fewer TSS than the other two classifiers, due to its high cutoffs (i.e. $2^{17.01}$ reads) to achieve FDR of 0.05. Number of TSS identified based on TIEScore and CAGE read count is comparable. TIEScore rescued 35,843 TSS discarded by the CAGE read count classifier cutoff at $2^{5.86}$ reads, implying its ability to identify lowly expressed TSS with acceptable FDR by synergizing information from CAGE and RNA-seq data.

3 | TIEScore Cutoffs and TSS Validation Rates



Supplementary Fig. 3 | TIEScore Cutoffs and TSS Validation Rates. a. Robust cutoff of TIEScore. The robust TIEScore cutoff (TIEScore=50.1) was based on the optimal cutoff determined from a ROC curve, with FDR=0.077. **b,** Permissive and stringent cuttofs of TIEScore. **c,** Cumulative percentage of TSS. '100%' refers all TSS in raw FANTOM CAT. **d,** FANTOM CAT TSS validation rate by DHS. Percentages of TSS at various TIEScore (bin=0.1) overlaps with DHS were calculated. Permutated TSS refers to randomly sampled genomic positions as control. **e,** TSS validation by RAMPAGE[14]. The percentages of TSS at various TIEScore (bin=0.1) that could be validated by RAMPAGE at various ranges were calculated and loess-smoothed curves were plotted. In **c-e, vertical dashed lines** represent the three TIEScore cutoffs as in **b**.

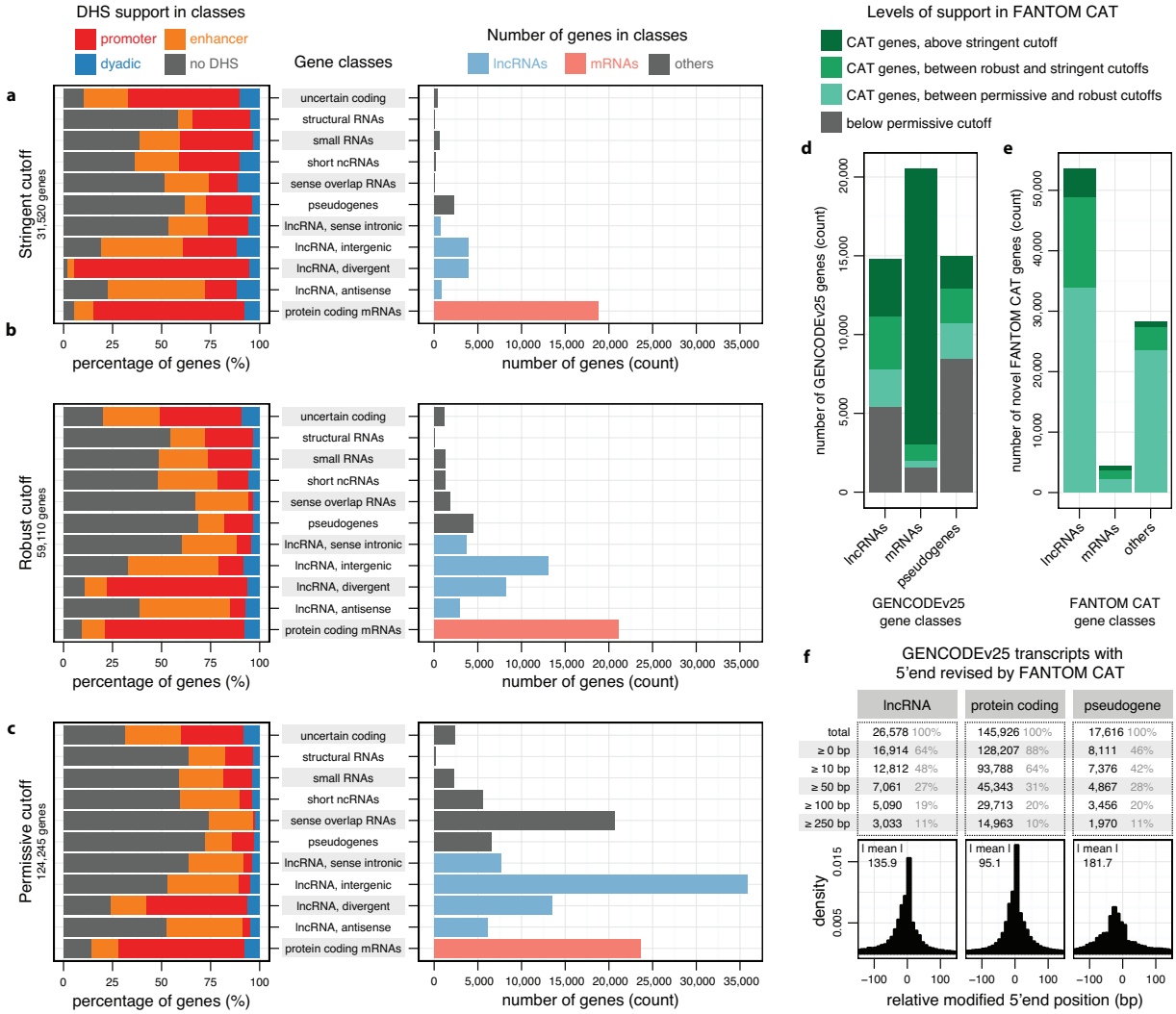
3.1 Definition of TIEScore cutoffs

Based on a ROC curve (**Supplementary Fig. 3a**), we derived an optimal TIEScore cutoff (TIEScore=50.1, referred to as robust cutoff), with ROC-AUC=0.89 and FDR=0.077. We also defined two additional values (**Supplementary Fig. 3b**), the permissive (TIEScore=35.3) and stringent (TIEScore=60.1) cutoffs, based on FDR chosen as three times more relaxed (FDR=0.231) and strict (FDR=0.026) than the robust FDR.

3.2 Properties at various TIEScore cutoffs

At the robust cutoff, 46.9% of all TSS were retained (**Supplementary Fig. 3c**). We observed high correlations between TIEScore and TSS validation rate by DHS (Pearson's r=0.94, Methods, **Supplementary Fig. 3d**) and by RAMPAGE[14] (mean Pearson's r=0.95, Methods, **Supplementary Fig. 3e**), suggesting that TIEScore quantitatively identifies genuine TSS.

4 | Definition of Genes, Coding Potential and Comparison with GENCODEv25



Supplementary Fig. 4 | Definition of FANTOM CAT Genes and Comparison with GENCODEv25. In **a-c**, **Left column**: Percentage of genes within each gene class supported by different types of Roadmap regulatory regions[7], based on overlapping of their strongest TSS with the Roadmap DHS[7]. At all cutoffs, the majority of protein coding mRNAs are supported by promoter DHS. **Right column**: Number of genes within each gene class. **a**, **b** and **c**, refer to FANTOM CAT genes at stringent, robust and permissive TIEScore cutoff respectively. **d**, GENCODEv25 genes in FANTOM CAT. The number of GENCODEv25 lncRNA, mRNA and pseudogene genes supported at various levels of FANTOM CAT was plotted. **e**, FANTOM CAT genes novel to GENCODEv25. A FANTOM CAT gene is novel to GENCODEv25 if none of its transcripts is 'compatible' with a GENCODEv25 transcript. **f**, Revision of 5'ends of GENCODEv25 transcripts. **Upper**: Number of GENCODEv25 transcripts revised by various extents. 'Total' refers to total number of GENCODEv25 transcripts. '≥*X* bp' refers to the number of GENCODEv25 transcripts revised by at least |*X*| bp. **Lower**: Distributions of the 5'end of GENCODEv25 lncRNA, mRNA and pseudogene transcripts relative to their 'compatible' FANTOM CAT transcripts (i.e. relative modified 5'end position) were plotted. The absolute mean within each class was indicated.

4.1 Definition FANTOM CAT genes

FANTOM CAT genes were defined based on clustering of transcript models in raw FANTOM CAT (after reducing complexity) using a custom perl script. Specifically, non-GENCODEv19 transcripts with exon boundaries within ±500bp to that of a GENCODEv19 transcript were first assigned to the corresponding GENECODEv19 genes. Single exon non-GENCODEv19 transcripts within ±500bp of a GENCODEv19 exon were also assigned to the corresponding GENECODEv19 gene. The non-GENCODEv19 transcripts spanning multiple GENECODEv19 genes were defined as chimeric transcripts and removed (as likely assembly artifacts). Remaining unassigned non-GENCODEv19 transcripts with exon boundaries within ±500bp were then recursively clustered, and all these resulting transcript clusters were referred to as novel genes outside GENCODEv19. Applying various TIEScore cutoffs to the raw FANTOM CAT, we defined 124,245 permissive, 59,110 robust and 31,520 stringent genes (**Supplementary Fig. 4a-c**). Gene classes of

FANTOM CAT genes assigned to GENCODEv19 protein coding genes, pseudogenes and small RNA genes were directly inherited from their biotypes[25]. Other gene classes in FANTOM CAT were defined as follows.

4.2 Definition of gene classes

Coding potential of all transcripts in raw FANTOM CAT was evaluated using the Coding Potential Assessment Tool (CPAT, version 1.2.232[15]) with default parameters on hg19. Transcripts with CPAT score <0.364 and no open reading frames (ORF) ≥300nt (based on getorf[16]) are defined as non-coding. A FANTOM CAT gene is defined as non-coding if all its transcripts are non-coding, or its GENCODEv19 biotype is annotated as non-coding[17]. A gene is defined as coding when at least 50% of its transcripts are coding and there is at least one transcript with an ORF ≥300nt. Otherwise the gene is classified as 'coding uncertain'. Non-coding genes generating at least one transcript with total exonic length ≥200nt are defined as lncRNA genes, and those <200nt are defined as 'short ncRNA'. LncRNAs are then classified in ascending order as follows: divergent lncRNAs, sense intronic lncRNAs, antisense lncRNAs and intergenic lncRNAs. Divergent lncRNAs: genes with its strongest CAGE cluster within ±2kb on the opposite strand of any CAGE clusters of GENCODEv19 protein coding genes or pseudogenes. Sense intronic lncRNA: lncRNA genes 1) initiating within the intron of another FANTOM CAT gene, 2) with at least 50% of their genic region overlapping with the genic region of any another genes, 3) with its strongest CAGE cluster not overlapping exons of other genes, and 4) containing ≥10 CAGE reads, or otherwise defined as 'other sense overlap RNA'. Antisense lncRNAs: genes with ≥50% of their genic region overlapping with the genic region of GENCODEv19 protein coding genes or pseudogenes on the opposite strand. Intergenic lncRNAs: the remaining lncRNA genes that could not be assigned to any of the above categories.

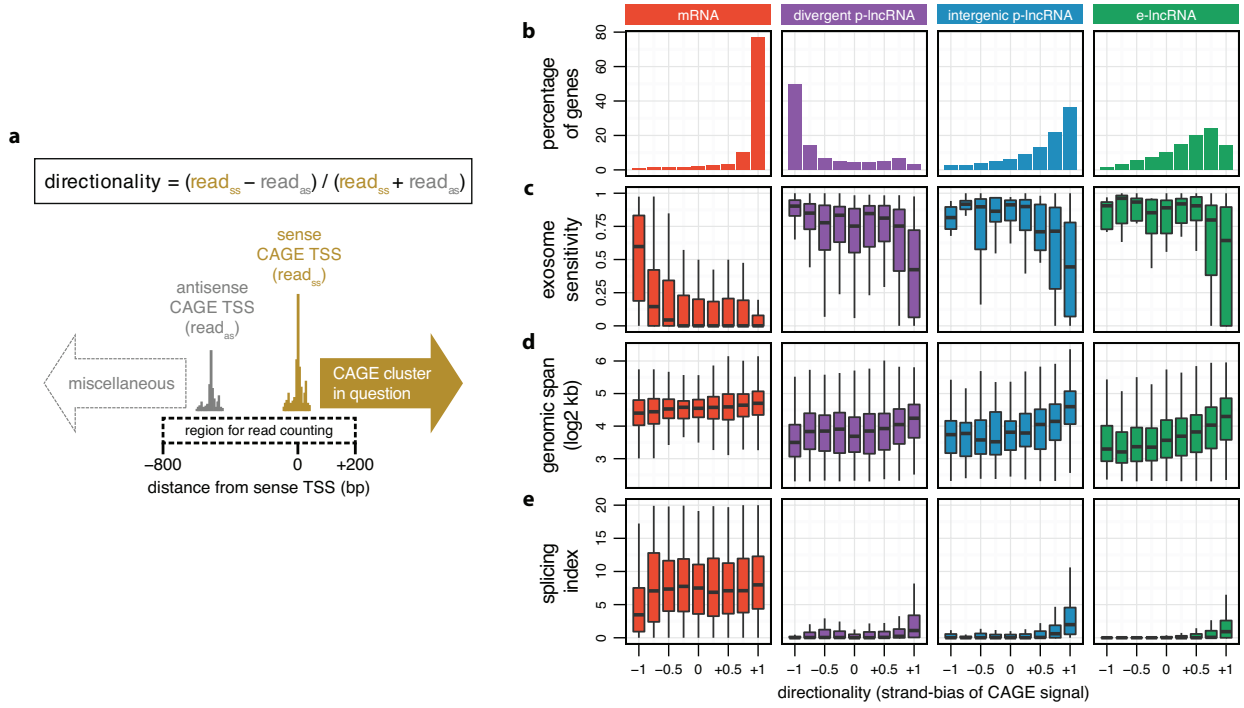
4.3 Comparison between FANTOM CAT and GENCODEv25

A GENCODEv25 gene is supported by FANTOM CAT if at least for one of its transcripts is 'compatible' with a FANTOM CAT transcript. A GENCODEv25 transcript is defined as compatible with FANTOM CAT transcripts if all of its exon boundaries are within ±500bp to that of a FANTOM CAT transcript, and *vice versa*. A FANTOM CAT gene is defined as novel if none of its transcripts is compatible to GENCODEv25 transcripts. About 33.86% (5,013 of 14,801) of the GENCODEv25 lncRNA genes (**grey stack**, **Supplementary Fig. 4d**) are below the permissive TIEScore cutoff. At the robust cutoff, FANTOM CAT covered 18,269 of the 20,132 protein-coding genes in GENCODEv25 (**Supplementary Fig. 4d**). The robust FANTOM CAT covered 6,994 of 14,801 lncRNA genes in GENCODEv25 (**Supplementary Fig. 4d**) and added 19,723 novel lncRNA genes outside of GENCODEv25 (**Supplementary Fig. 4e**). In addition, the 5'ends of 16,914 lncRNA transcripts in GENCODEv25 were revised using FANTOM CAT models, which resulted in their improved accuracy by a mean of 135.9bp (**Supplementary Fig. 4f**).

4.4 Annotation of open reading frames in FANTOM CAT.

As some previously annotated putative lncRNAs transcripts have been shown to associate with ribosomes and thus may encode for short peptides[18], we further annotated the coding potential of open reading frames, but were not used for defining the coding status of a FANTOM CAT gene. Coordinates of ORFs on all FANTOM CAT transcripts (n=861,584) were extracted using getorf[16], with minimum 30nt requiring both start and stop codons. Only the top 5 longest ORFs per transcript were retained, which yielded 3,006,858 non-redundant ORFs. Coordinates of the ORFs were converted from transcript level to genomic level. A pre-computed 46-way whole genome alignment[19] of hg19 was downloaded from UCSC Genome Browser[20]. Only 27 species were retained in the alignment based on its intersection with the 58-mammal phylogeny used in PhyloCSF[21], including Alpaca, Armadillo, Baboon, Bushbaby, Cat, Chimp, Cow, Dog, Dolphin, Elephant, Gorilla, Guinea pig, Hedgehog, Horse, Human, Marmoset, Megabat, Microbat, Mouse, Orangutan, Pika, Rabbit, Rat, Rhesus, Shrew, Squirrel and Tenrec. Interspecies alignments of the ORF were extracted using mafsInRegion from UCSC Genome Browser[20]. Coding potential of these ORFs were assessed using PhyloCSF[21] and RNAcode[22] using default settings, based on the phylogenetic information from the interspecies alignments. We further identified potentially translated small ORFs (sORF) based on ribosome profiling data in sorfs.org[23]. An ORF is defined as coding if its score ≥41 in PhyloCSF[21], P<0.001 in RNAcode[22] or it matches an sORF with 'good' floss-score as defined in sorfs.org[23]. The annotations of all ORFs in FANTOM CAT are available at http://fantom.gsc.riken.jp/cat/. Note: Of 27,919 lncRNAs genes in robust FANTOM CAT, 7,722 of them were annotated as lncRNAs in GENCODEv19[24] and the remaining 20,197 non-GENCODEv19 lncRNA genes where all their transcripts lacked coding potential (CPAT score <0.364 and no ORF ≥300nt). Of these 20,197 non-GENCODEv19 lncRNA genes, only 1,087 had putative ORFs with additional evidence of coding potential (in either PhyloCSF[21], RNAcode[22] or sorfs.org[23]). We provide these annotations on the web resource (http://fantom.gsc.riken.jp/cat/), yet we still consider these as genuine lncRNAs for further analysis.

5 | Directionality, Exosome Sensitivity and Transcript Properties



Supplementary Fig. 5 | Directionality, Exosome Sensitivity and Transcript Properties. **a**, Definition of directionality. 'Miscellaneous' refers to any antisense TSS (if any). Directionality of a CAGE cluster is defined as: ($read_{SS}$–$read_{AS}$)/($read_{SS}$+$read_{AS}$), where $read_{SS}$ and $read_{AS}$ are the number of CAGE read counts (across the 1,897 FANTOM5 samples) within the –800 to +200nt of its prominent TSS on the sense and antisense strand, respectively. **b** the percentages of genes within each category for each bin of directionality. **c**, Exosome sensitivity is measured as the relative fraction of CAGE signal observed after exosome knockdown in HeLa-S3 cells[25]. **d**, genomic span, defined as the length between the 5' most to 3' most base of a gene. **e**, splicing index, calculated as the average of the sum of the ratio of junction spanning reads to spliced reads of each of its introns across 107 RNA-seq libraries. The box plots show the median, quartiles and Tukey whiskers of the measurements for genes at each bin of directionality.

5.1 Rationale

Transcription initiation is intrinsically bidirectional[26] (**Supplementary Fig. 5a**). Functionally distinct RNA species were previously[25] categorized by their transcriptional directionality (strand-bias of transcription initiation, referred to as directionality) and by exosome-sensitivity (sensitivity to the ribonucleolytic RNA exosome complex). For each lncRNA category we examined the relationship between these features and the length and splicing of the observed transcripts (**Supplementary Fig. 5b-e**).

5.2 Calculation of directionality, splicing index, genomic span and exosome sensitivity

Directionality of a given CAGE cluster is measured as the bias of CAGE signal on the sense versus antisense strand (relative to the TSS), as described[25] with modifications as follows. A zero value implies perfectly balanced CAGE signal on both strands, and values of +1 and −1 implies strongly biased CAGE signal towards the sense or antisense strand, respectively. Directionality of a CAGE cluster is defined as follows: ($read_{SS}$–$read_{AS}$)/($read_{SS}$+$read_{AS}$), where $read_{SS}$ and $read_{AS}$ are the number of read counts from all FANTOM5 CAGE datasets (n=1,897, **Supplementary Table 1**) within the –800 to +200nt of its prominent TSS on the sense and antisense strand, respectively (**Supplementary Fig. 5a**). Directionality of a gene is defined by the directionality of its strongest CAGE cluster. Splicing index of a gene is defined as the average of the sum of the ratio of junction spanning reads to spliced reads of each of its introns across 107 RNA-seq libraries (libraries mentioned in Methods). Genomic span of a gene is defined as the length between its 5' most to 3' most genomic location. Exosome sensitivity of a CAGE cluster is measured as the relative fraction of CAGE signal observed after exosome knockdown in HeLa-S3 cells as previously described[25].

5.2 Directionality p-lncRNAs

Consistent with previous findings[25], transcription initiation regions (TIRs) that are predominantly transcribed from the sense strand (directionality ≈+1, in all four categories, **Supplementary Fig. 5b**) produce less exosome-sensitive (**Supplementary Fig. 5c**), generally longer (**Supplementary Fig. 5d**) and more spliced and RNAs (**Supplementary Fig. 5e**). Therefore, directionality of TIR somewhat reflects the properties of its produced RNAs. As expected, the TIRs of mRNAs predominantly generate sense transcripts (directionality ≈+1, **red**) and are only mildly exosome sensitive,

whereas the p-lncRNAs generated from divergent antisense transcription from these mRNA promoters are PROMPT[27] like (directionality ≈–1, **purple**), largely exosome sensitive, short and rarely spliced. In contrast intergenic p-lncRNA TIRs are mostly biased towards generating sense transcripts (directionality >0, **blue**). In addition, ~40% of intergenic p-lncRNAs predominantly generating sense transcripts (directionality ≈+1, **blue**), such as *MALAT1*[28], are relatively resistant to exosome degradation.

5.3 Directionality e-lncRNAs

About 76% of e-lncRNAs arose from balanced bidirectional TIRs (directionality ≥−0.8 and ≤+0.8, **green**), while about 22% were generated from unidirectional TIRs (directionality >+0.8, **green**). These unidirectional TIR derived e-lncRNAs are generally less exosome sensitive, longer, more spliced (directionality >+0.8, **green**). This study therefore expand our previous catalog of bidirectionally transcribed enhancer regions[8] by including unidirectional e-lncRNAs. These contain previously identified functional e-lncRNA such as *CCAT1*, an unidirectionally transcribed from a super-enhancer, which promotes long-range chromatin looping and regulates MYC transcription[29].

6 | Sequence Features at LncRNA TSS Supported by Different DHS Types

Supplementary Fig. 6 | Sequence Features at LncRNA TSS Supported by Different DHS Types. **Black dashed lines**: whole genome background. **Solid lines**: TSS of lncRNA genes and control regions randomly sampled from whole (whole genome) or unannotated portion of the genome (unannot. regions). Colors of the lines correspond to the type of DHS support as indicated on top. **a**, Epigenomic features surrounding TSS. Distributions of Roadmap DNase-seq and ChIP-seq (Methods) were plotted. **b**, Genomic features surrounding TSS. Distributions of rejected substitution (RS) score[30], CpG island, polyadenylation site signal (PAS) and 5' splicing site (5'SS), were plotted. **c**, Core promoter motifs. Distributions of TATA box and initiator (INR) motif, were plotted.

.

6.1 Genuineness TSS of lncRNA gene classes supported by different DHS types

Despite the relatively weaker selective constraints of e-lncRNAs and intergenic p-lncRNAs TIR as measured using RS scores[30], we observed an enrichment of sequence features associated with regulated transcription initiation, including TATA-box, INR and CpG island (absent for e-lncRNAs) as well as signals conducive to generating long transcripts including 5' SS enrichment and PAS depletion. These sequence features suggest that at least a subset of these TIR have undergone selection for both transcription initiation and elongation. For the remaining lncRNAs lacking DHS support (**grey**), we still observed noticeable enrichment of 5' SS, INR and TATA-box motifs and depletion of PAS, suggesting that a subset of these TSS is likely to be genuine despite the lack of DHS support.

B | Online Resources

All resources mentioned below are available at http://fantom.gsc.riken.jp/cat/.

1 | Assembly, Expression Atlas and other Resources

1.1 Assembly: Cutoffs and identifiers

We provide The FANTOM CAT assembly at 4 cutoffs (as *.gtf),

- ***Raw***: without TIEScore cutoff, FDR=0.285, contains 100% of all TSS
- ***Permissive***: TIEScore ≥35.3, FDR=0.231, contains 88.6% of all TSS
- ***Robust***: [**optimal, recommend to use**] TIEScore ≥50.1, FDR=0.077, contains 46.9% of all TSS
- ***Stringent***: TIEScore ≥60.1, FDR=0.026, contains 19.4% of all TSS

At each of the 4 TIEScore cutoffs, the assembly consists of 3 types of basic items (as *.bed), 1) genes, 2) transcripts and 3) CAGE clusters. Each of these items has a unique identifier and associates to items of the other types as specified in ID mapping tables (as *.txt):

- ONE transcript associates with ONLY ONE CAGE cluster;
- ONE transcript associates with ONLY ONE gene;
- ONE CAGE cluster can associate with multiple transcripts;
- ONE CAGE cluster can therefore associate with multiple genes;

Gene identifiers (mainly inherited from GENCODEv19):

- ***ENSG*.***: inherited from GENCODEv19, e.g. ENSG00000141577.9;
- ***CATG*.***: novel genes from FANTOM CAT, e.g. CATG00000015547.1;

Transcript identifiers (mainly inherited from GENCODEv19):

- ***ENST*.***: inherited from GENCODEv19[17], e.g. ENST00000571292.1;
- ***ENCT*.***: novel transcripts from ENCODE[6] transcript models, e.g. ENCT00000114868.1;
- ***HBMT*.***: novel transcripts from Human BodyMap v2.0[4] transcript models, e.g. HBMT00000386110.1;
- ***MICT *.***: novel transcripts from miTranscriptome[5] transcript models, e.g. MICT00000096911.1;
- ***FTMT*.***: novel transcripts from FANTOM5 RNA-seq transcript models, e.g. FTMT25100014196.1;

CAGE Cluster identifiers (ALL inherited from FANTOM5 'DPI CAGE Clusters'):

- chr*:*..*,*: as chromosome:start..end,strand, inherited from FANTOM5, e.g. chr1:121258327..121258338,+;

1.2 Expression Atlas

At each of the 4 TIEScore cutoffs, we provide the following expression tables (1,897 FANTOM5 CAGE libraries):

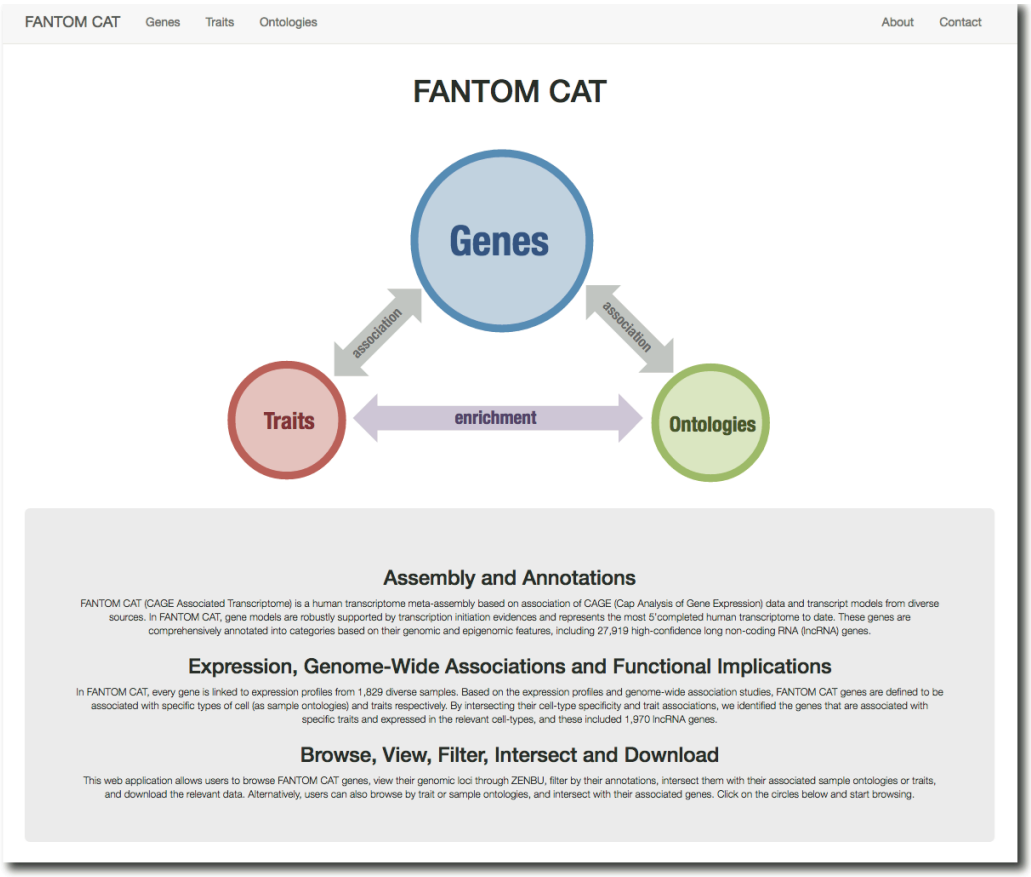
- ***tag count per CAGE cluster***: CAGE tags count within the flanking region (±50nt) of the CAGE cluster;
- ***tag count per gene***: sum of CAGE tags of CAGE clusters associated with the gene;
- ***rle cpm per CAGE cluster***: rle normalized (by edgeR[31]) cpm (count per millions) the CAGE cluster;
- ***rle cpm per gene***: sum of CAGE tags of CAGE clusters associated with the gene;

1.3 Other Resources

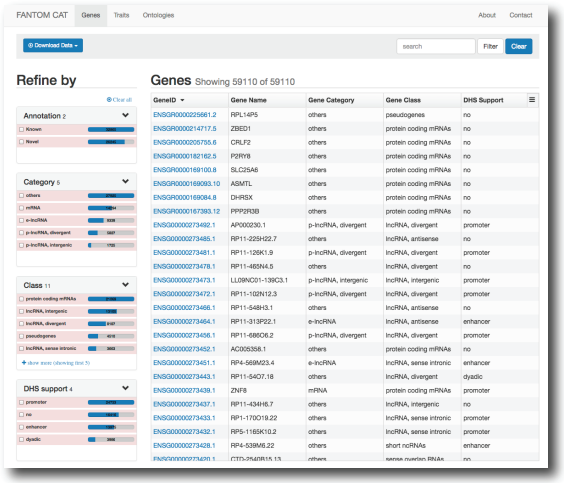
We also provide the following resources for the FANTOM CAT as robust TIEScore cutoff:

- ***Differential expression***: differential expression analysis (edgeR[31]) of 25 manually curated series;
- ***Sample ontology association***: gene-based association with FANTOM5 sample ontology terms;
- ***Trait association***: Gene-based association with traits in GWASdb[32] and PICS[33];
- ***Expression specificity***: Gene-based expression levels and specificity in primary cell facets;
- ***eQTL Co-expression***: Co-expression of lncRNA–mRNA pairs linked by GTEx[34] eQTL-associated SNPs;
- ***Selective constraints***: RS score and GERP elements[30] at TIR and exon;
- ***Orthologous TSS activities***: Orthologous TSS activities in mouse, rat, dog and chicken;
- ***ORF Annotations***: coding potential of all ORF based on phyloCSF[21], RNACode[22] and sorf.org[23];
- ***Transposable elements***: TIR association with transposable elements;
- ***Directionality***: Transcriptional directionality of TIR based on 1,897 FANTOM5 CAGE libraries;
- ***Exosome Sensitivity***: Exosome sensitivity based on exosome knockdown in HeLa cells[25];
- ***Splicing Index***: Gene-based splicing index in 107 RNA-seq libraries;

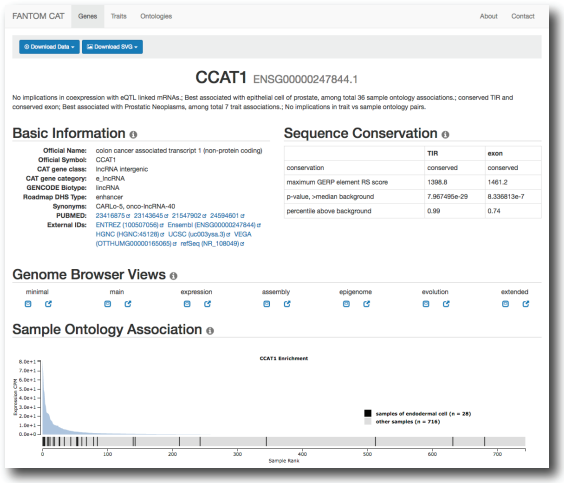
2 | Overview of FANTOM CAT Browser



Supplementary Fig. 7 | Landing Page of FANTOM CAT Browser. The FANTOM CAT Browser is hosted at http://fantom.gsc.riken.jp/cat/v1/#/. This web application allows users to browse FANTOM CAT genes, view their genomic loci through ZENBU[35], filter by their annotations, intersect them with their associated sample ontologies or traits, and download the relevant data. Alternatively, users can also browse by trait or sample ontologies, and intersect with their associated genes. Click on the circles below and start browsing Genes, Traits and Ontologies.



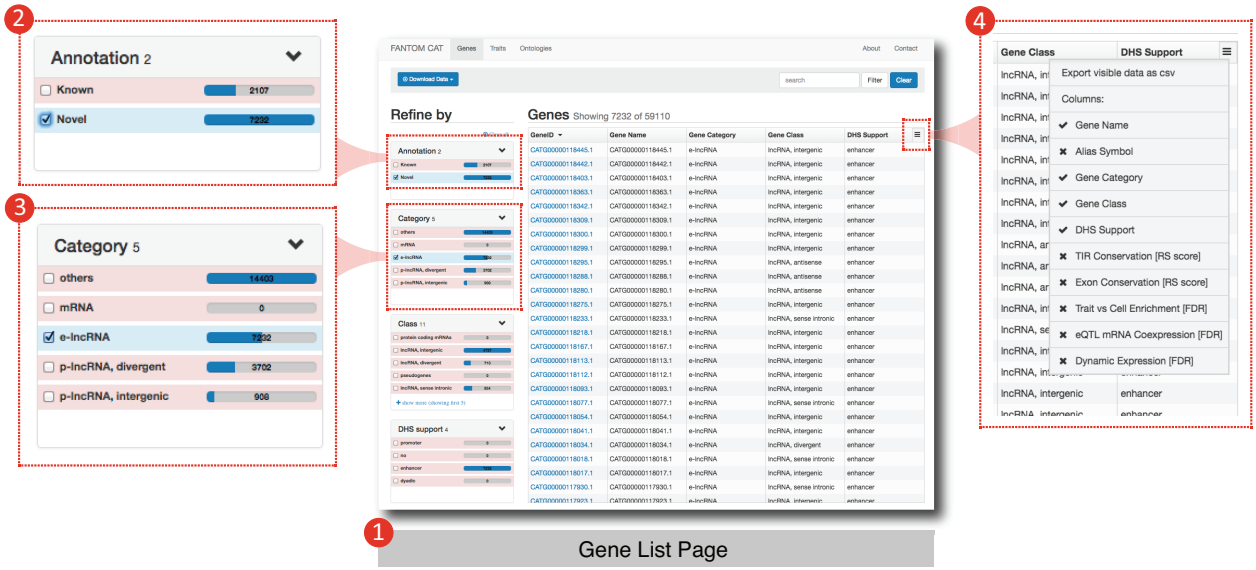
Gene List Page



Gene Individual Page

Supplementary Fig. 8 | Browsing Genes from the Gene List and Individual Page. On the Gene List Page (http://fantom.gsc.riken.jp/cat/v1/#/genes), users can select for genes based on the filters on the left if the list. Click on the GeneID on the list to browse an individual gene. The Gene Individual Page displays the basic information of the gene and its associated sample ontologies and traits. User can navigate the between genes, sample ontologies and traits through the links. Alternatively, users can start by browsing the list of sample ontologies and traits and reach their associated genes.

3 | Use Case 1: Download a List of Novel e-lncRNAs

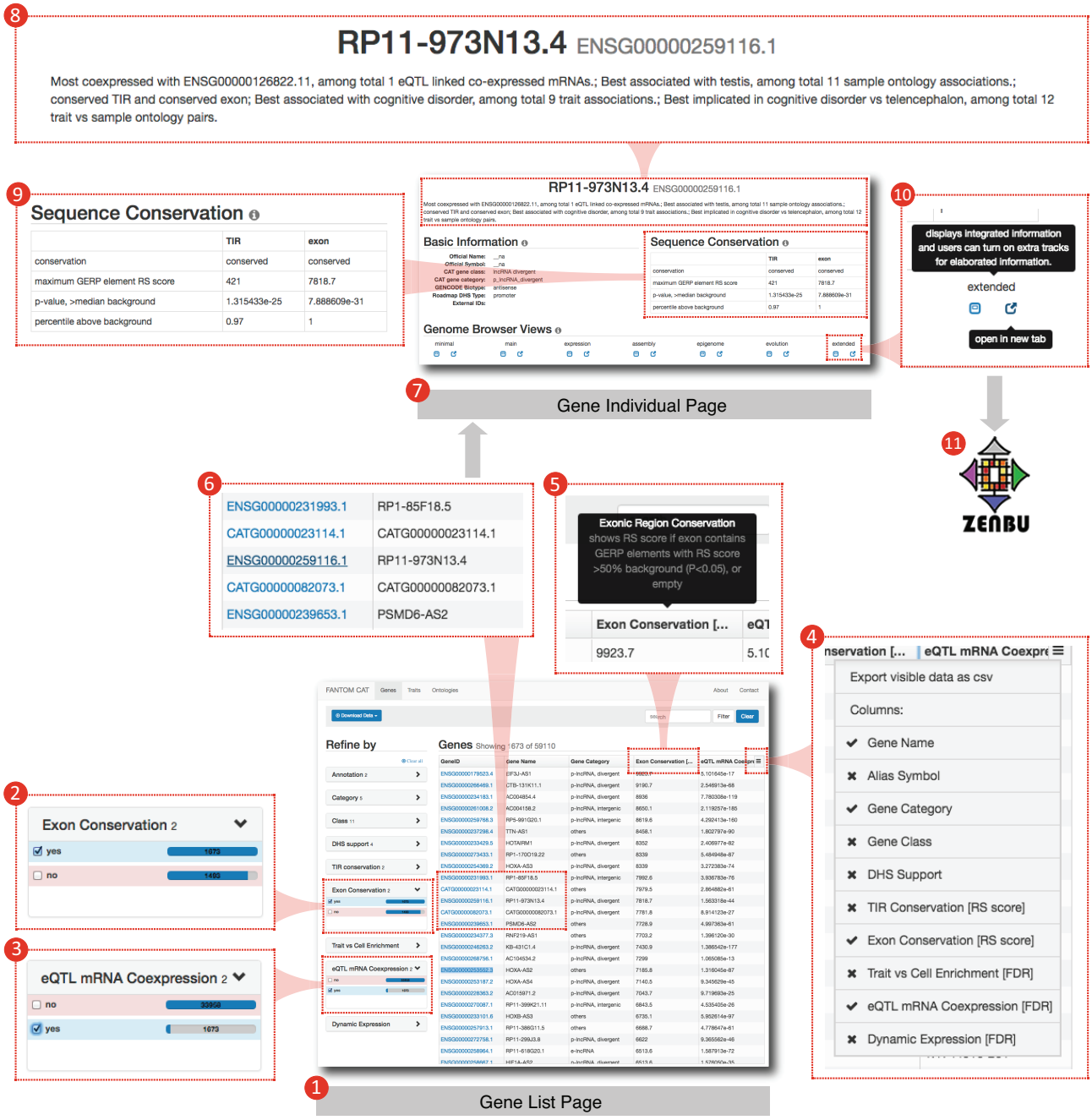


Supplementary Fig. 9 | Download a List of Novel e-lncRNAs.

Steps:

1. Go to Gene List Page by clicking '**Genes'** in the Landing Page;
2. Check '**novel**' in the filter '**Annotations**' to filter for genes that are not annotated in GENCODEv19;
3. Check '**e-lncRNA**' in the filter '**Category**', and this should return 7,232 genes;
4. Choose the data columns to be displayed (and thus downloaded) and click '**Export visible data as csv**';

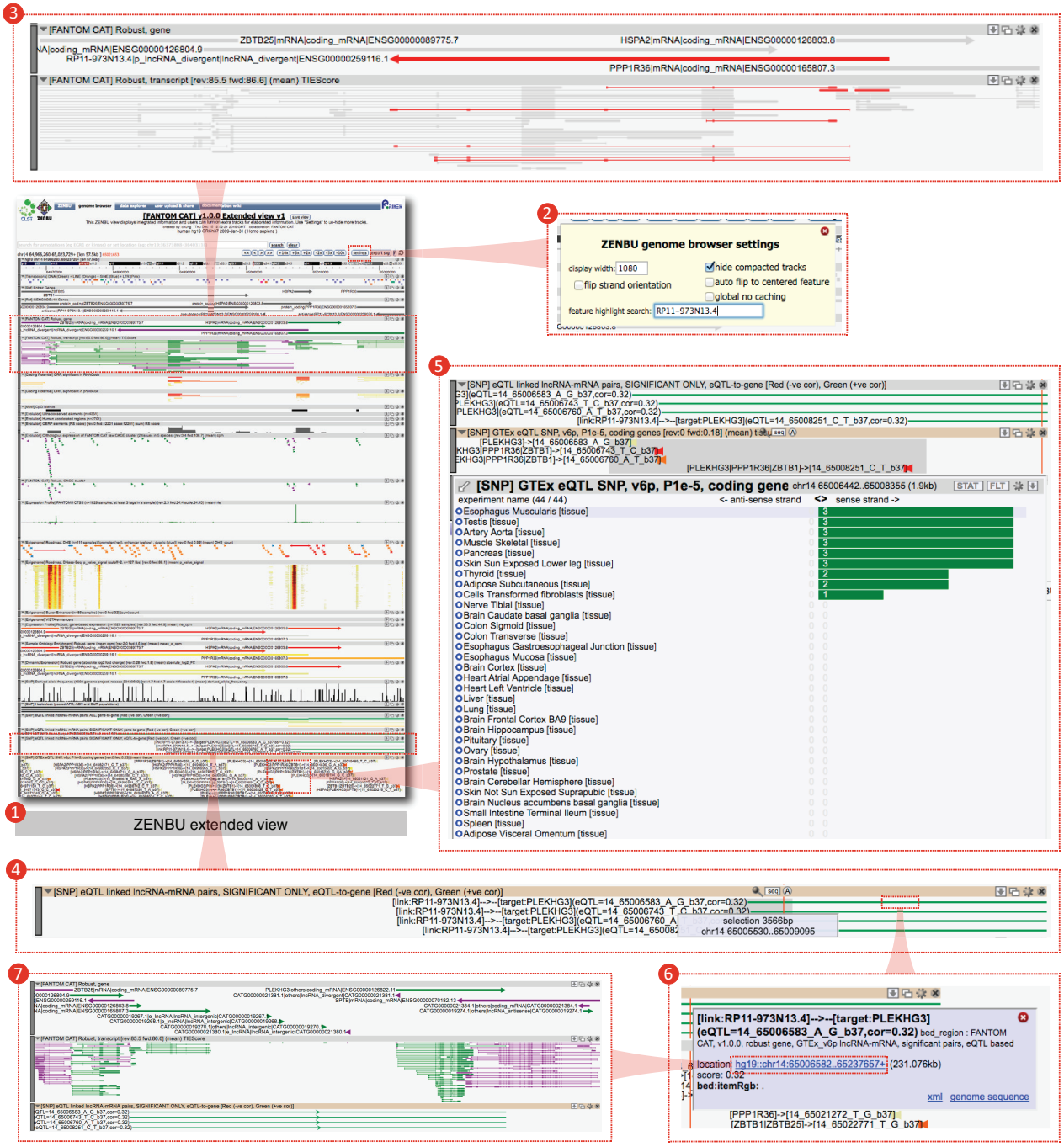
4 | Use Case 2: Explore LncRNAs with Conserved Exons and Implicated in eQTL



Supplementary Fig. 10 | Explore LncRNAs with Conserved Exons and Implicated in eQTL.

Steps:

1. Go to Gene List Page by clicking '**Genes**' in the Landing Page;
2. Check 'yes' in the filter '**Exon Conservation**' to filter for genes with conserved exons;
3. Check 'yes' in the filter '**eQTL mRNA Coexpression**', and this should return 1,673 genes;
4. Choose the data columns to '**Exon conservation**' and '**eQTL mRNA Coexpression**' be displayed;
5. Click on header of '**Exon conservation**' column to sort genes by level of conservation at their exons;
6. Click on one of the genes on the top (i.e. highly conserved exon), e.g. ENSG00000259116.1;
7. It should reach the individual page for gene ENSG00000259116.1;
8. Examining the summary for the gene, it notes ENSG00000259116.1 is coexpressed with eQTL-linked ENSG00000126822.11;
9. Examining the conservation summary of the gene, it notes exon of ENSG00000259116.1 is conserved;
10. Click on the '**extended**' to view the ENSG00000259116.1 loci;
11. *Continue next page*: in extended ZENBU view, it shows the eQTL information at the locus;

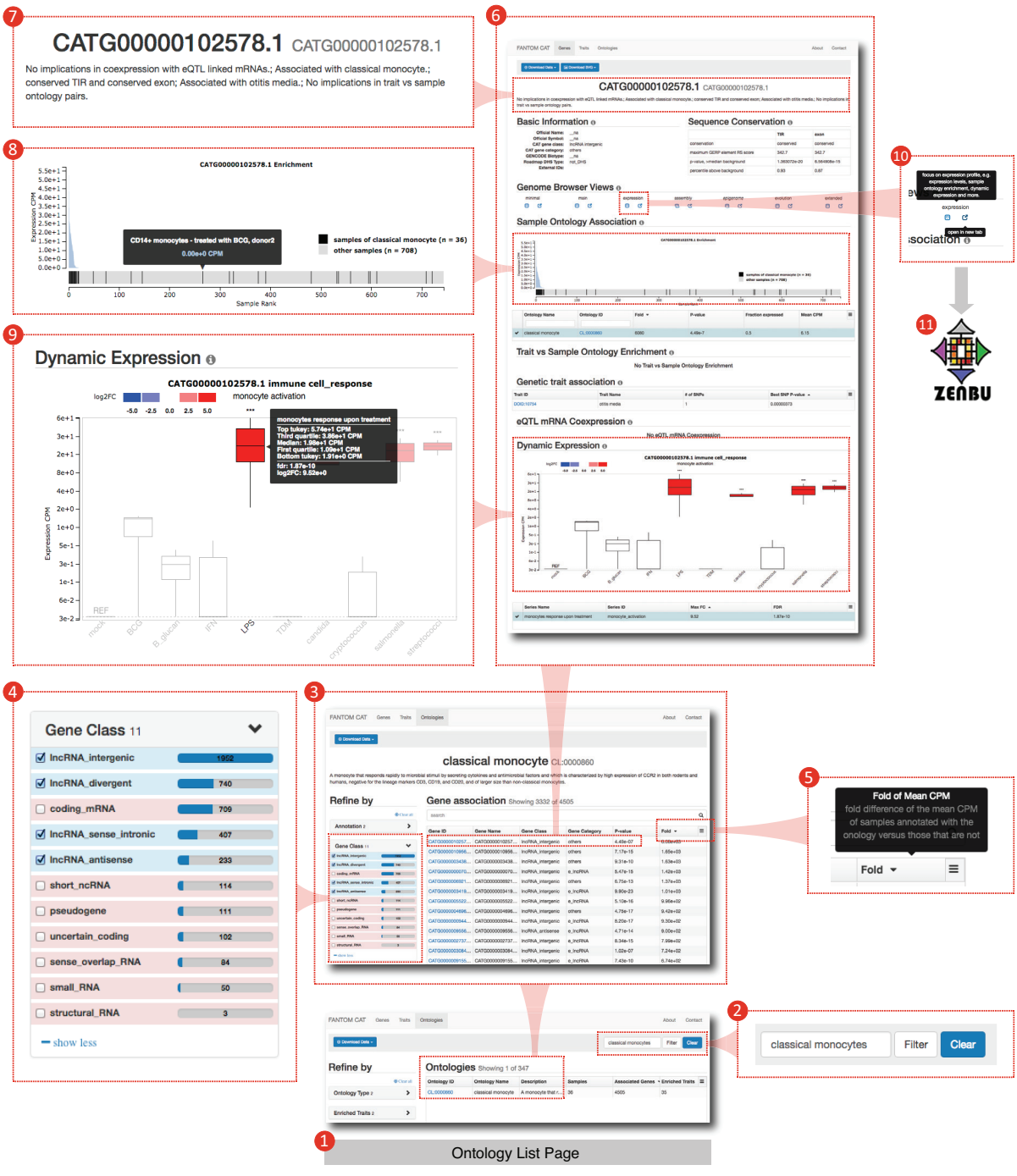


Supplementary Fig. 11 | Visualize an LncRNA Implicated in eQTL in ZENBU.

Steps:

1. Continuing from the end of Supplementary Fig. 10, you should see the '**ZENBU extended view**' for ENSG00000259116.1;
2. Click '**settings**' and in '**feature highlight search**' input 'RP11-973N13.4' (gene name of ENSG00000259116.1) to highlight the locus of RP11-973N13.4;
3. Examine the red highlighted gene and transcript of RP11-973N13.4. To cancel the highlight, repeat step 2 and delete the search string in the setting dialogue box;
4. Click on track '**[SNP] eQTL linked lncRNA-mRNA pairs, SIGNIFICANT ONLY, eQTL-to-gene**', which shows connections between the eQTL SNPs to the target coexpressed mRNA, highlight the regions flanking the 4 eQTL SNPs (i.e. left ends of the 4 green lines) and click the '**magnifying glass**' icon to zoom in;
5. In the zoom in view, click on track '**[SNP] GTEx eQTL SNP, v6p, P1e-5, coding genes**', and highlight the 4 eQTL SNPs, and bring up the display panel (at the bottom) showing the active tissues of the highlighted eQTL SNPs;
6. To zoom out to view the lncRNA (i.e. RP11-973N13.4) and the linked mRNA (i.e. PLEKHG3), click on one of the connections in the track '**[SNP] eQTL linked lncRNA-mRNA pairs, SIGNIFICANT ONLY, eQTL-to-gene**', then click the range to zoom out;
7. A view contains the coexpressed lncRNA–mRNA pair linked by 4 eQTL SNP, you can rearrange the tracks by dragging the title bars as desired, e.g. move up the track '**[SNP] eQTL linked lncRNA-mRNA pairs, SIGNIFICANT ONLY, eQTL-to-gene**';

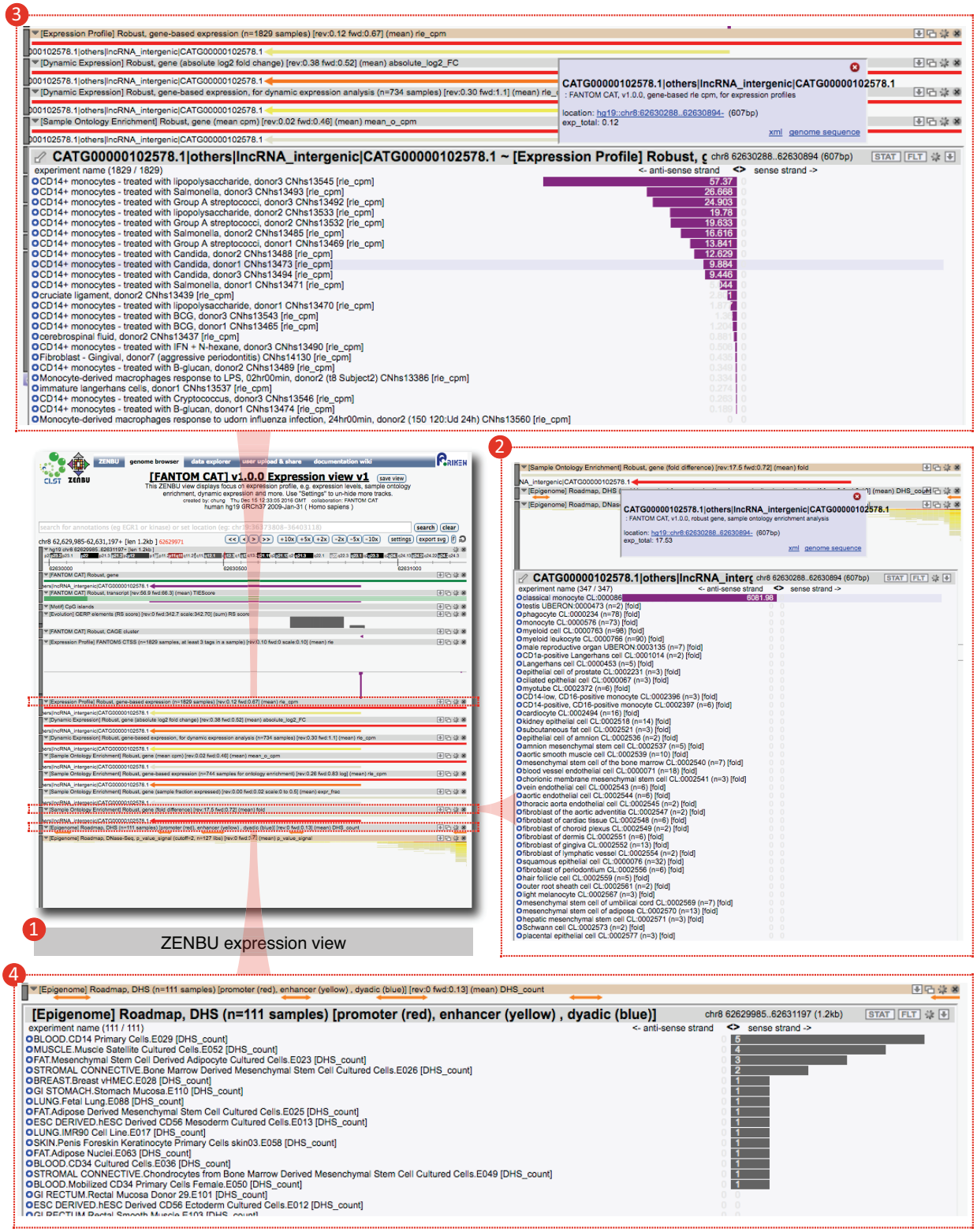
5 | Use Case 3: Explore LncRNAs Enriched in Classical Monocytes



Supplementary Fig. 12 | Explore LncRNAs Enriched in Classical Monocytes.

Steps:

1. Go to Ontology List Page by clicking '**Ontologies**' in the Landing Page;
2. Type '**classical monocytes**' and press '**filter**';
3. Click on '**CL:0000860**' to enter the ontology individual page for classical monocytes;
4. Check the 4 lncRNA classes in the filter '**Gene Class**' to filter for lncRNA genes;
5. Click on header of '**Fold**' column to sort genes by level of enrichment in classical monocytes;
6. Click on the genes on the top (i.e. most enriched in classical monocytes), i.e. CATG00000102578.1;
7. Examining the summary for the gene, it notes CATG00000102578.1 is associated with classical monocytes;
8. Hover over the enrichment plot to examine the expression of CATG00000102578.1 in individual samples of classical monocytes;
9. Hover over the box plot to examine the dynamic expression of CATG00000102578.1 in classical monocytes stimulation;
10. Click on the '**expression**' to view the CATG00000102578.1 loci;
11. *Continue next page*: in expression ZENBU view, it shows the expression information at the locus;

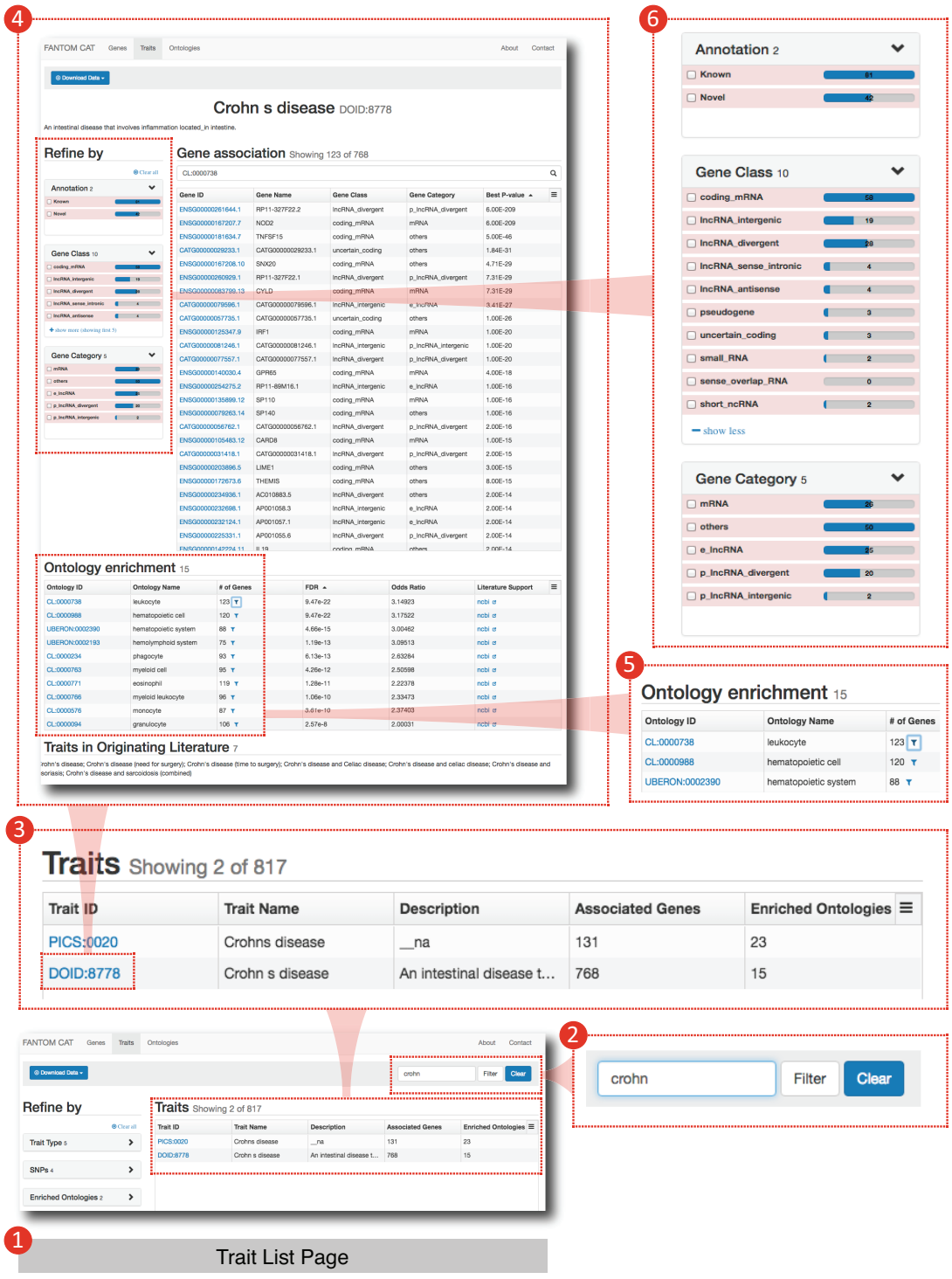


Supplementary Fig. 13 | ZENBU Expression View of an LncRNA Enriched in Classical Monocytes.

Steps:

1. Continuing from the end of Supplementary Fig. 12, you should see the '**ZENBU expression view**' for CATG00000102578.1;
2. Click on CATG00000102578.1 in track '**[Sample Ontology Enrichment] Robust, gene (fold difference)**' and bring up the experiment panel (at the bottom) showing the fold enrichment of CATG00000102578.1 in classical monocytes;
3. Click on CATG00000102578.1 in track '**[Expression Profile] Robust, gene-based expression (n=1829 samples)**' and bring up the experiment panel (at the bottom) to show the expression of CATG00000102578.1 in 1,829 FANTOM5 samples;
4. Click on track '**[Epigenome] Roadmap, DHS (n=111 samples)**' to shows the tissues of the roadmap DHS are active within the displayed loci;

6 | Use Case 4: Explore Cell Types Associated with Crohn's Diseases



Supplementary Fig. 14 | Explore Cell Types Associated with Crohn's Diseases.

Steps:

1. Go to Trait List Page by clicking '**Ontologies**' in the Landing Page;
2. Type '**crohn**' and click '**filter'**;
3. Click on '**DOID:8778**' to enter the trait individual page for Crohn's Disease GWAS;
4. Reached trait individual page for Crohn's Disease GWAS;
5. Browse the cell-types that are associated Crohn's Disease in the '**Ontology Enrichment**' table, and click on the '**filter**' to filter for genes that are associated with Crohn's Disease and enriched in '**Leukocytes**', this should return 123 genes;
6. Examine the composition of the filtered genes, and go to individual gene page for gene based information;

C | Supplementary Tables

All Supplementary tables are available in the online version of the paper in Excel format.

1 | CAGE Library Information

Information of all FANTOM5 CAGE libraries ([human=1897]; [mouse=6]; [rat=6]; [dog=6]; [chicken=6]) used in this study.

2 | RNA-seq Library Information

Information of all FANTOM5 RNA-seq libraries (n=70) used in this study.

3 | FANTOM CAT Gene Information

Classification and essential information of FANTOM CAT robust genes (n=59110).

4 | Directionality, Exosome Sensitivity and Transcript Properties

Transcription directionality, exosome sensitivity, genomic span and splicing extent of FANTOM CAT robust genes (n=59110).

5 | FANTOM CAT Genes in LncRNAdb

A list of lncRNAdb matched FANTOM CAT lncRNAs (n=81).

6 | Conservation of TIR and Exon

Conservation measurements of transcription initiation region and exonic regions of FANTOM CAT robust genes (n=59110).

7 | Transposons at TIR

Overlap between various types of transposons at the TIR of FANTOM CAT robust genes (n=59110).

8 | Orthologous Transcription

Orthologous transcription activity of FANTOM CAT robust genes (n=59110) in 2 matched cell types of 4 other species.

9 | Expression Levels and Specificity in Primary Cell Facets

Expression levels and specificity of FANTOM CAT robust genes (n=59110) in primary cell facets (n=69).

10 | Sample Ontology Information

Information of the sample ontology terms (n=347) describing the selected FANTOM5 primary cell and tissue samples (n=744).

11 | Gene Association with Cell Types

Association between FANTOM CAT robust genes (n=59110) and sample ontology terms (i.e. cell types).

12 | Trait Information

Information of the trait terms (n=816) from GWASdb (GWAS SNPs) and PICS (fine-mapped SNPs).

13 | Gene Association with Traits

Association between FANTOM CAT robust genes (n=59110) and trait terms (i.e. traits).

14 | Curation of Cell type and Trait Pairs

Biological plausibility (based on manual literature curation) of significantly associated pairs of cell types and traits, and random control pairs.

15 | Genes Involved in Cell Type and Trait pairs

A list of FANTOM CAT robust genes involved in significantly associated pairs of cell-types and traits.

16 | eQTL-linked lncRNA and mRNA Pairs

A list of eQTL linked (GTEx V6p) lncRNA-mRNA pairs and their expression correlation in 1829 FANTOM5 samples.

17 | Gene-based Functional Evidence

A summary of functional evidences (TIR & exon conservation, trait implications and eQTL implications) of FANTOM CAT robust genes (n=59110).

18 | Grouping of Samples for Differential Expression

Information of FANTOM5 sample combinations for investigating dynamic differential expression.

19 | Differential Expression Results

Results dynamic differential expression of 25 series of FANTOM CAT robust genes

D | References

1. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–84 (2013).
2. Hestand, M. S. *et al.* Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. *Nucleic Acids Res.* **38,** e165 (2010).
3. Grau, J., Grosse, I. & Keilwagen, J. PRROC: Computing and visualizing Precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31,** 2595–2597 (2015).
4. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–1927 (2011).
5. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47,** 199–208 (2015).
6. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
7. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
8. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–61 (2014).
9. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).
10. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347,** 1010–4 (2015).
11. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9,** 215–6 (2012).
12. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32,** 462–464 (2014).
13. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: Visualizing classifier performance in R. *Bioinformatics* **21,** 3940–3941 (2005).
14. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23,** 169–180 (2013).
15. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41,** e74 (2013).
16. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
17. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
18. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *Elife* **3,** e03523 (2014).
19. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
20. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
21. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** i275--82 (2011).
22. Washietl, S. *et al.* RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17,** 578–94 (2011).
23. Olexiouk, V. *et al.* SORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **44,** D324–D329 (2016).
24. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22,** 1760–1774 (2012).
25. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* **5,** 5336 (2014).
26. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–20 (2014).
27. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322,** 1851–4 (2008).
28. Schmidt, L. H. *et al.* The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. *J. Thorac. Oncol.* **6,** 1984–92 (2011).
29. Xiang, J.-F. *et al.* Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res.* **24,** 513–531 (2014).
30. Davydov, E. V *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).
31. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
32. Li, M. J. *et al.* GWASdb v2: An update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44,** D869–D876 (2016).
33. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).
34. GTEx Consortium, Gte. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–60 (2015).
35. Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat. Biotechnol.* **32,** 217–219 (2014).