

The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity

Fairlie Reese^{1,2}, Brian Williams³, Gabriela Balderrama-Gutierrez^{1,2}, Dana Wyman², Muhammed Hasan Çelik², Elisabeth Rebboah², Narges Rezaie², Diane Trout³, Milad Razavi-Mohseni^{4,5}, Yunzhe Jiang^{6,7}, Beatrice Borsari^{6,7,8}, Samuel Morabito², Heidi Yahan Liang², Cassandra J. McGill^{1,2}, Sorena Rahamanian², Jasmine Sakr^{2,9}, Shan Jiang^{1,2}, Weihua Zeng^{1,2}, Klebea Carvalho², Annika K. Weimer¹⁰, Louise A. Dionne¹¹, Ariel McShane^{12,13}, Karan Bedi^{14,15}, Shaimae I. Elhajjajy¹⁶, Sean Upchurch³, Jennifer Jou¹⁰, Ingrid Youngworth¹⁰, Idan Gabdank¹⁰, Paul Sud¹⁰, Otto Jolanki¹⁰, J. Seth Strattan¹⁰, Meenakshi S. Kagda¹⁰, Michael P. Snyder¹⁰, Ben C. Hitz¹⁰, Jill E. Moore¹⁶, Zhiping Weng¹⁶, David Bennett^{18,19}, Laura Reinholdt¹¹, Mats Ljungman^{15,20}, Michael A. Beer^{4,5}, Mark B. Gerstein^{6,7,21,22,23}, Lior Pachter^{3,24}, Roderic Guigó^{8,25}, Barbara J. Wold³, and Ali Mortazavi^{1,2}

¹Developmental and Cell Biology, University of California, Irvine, Irvine, USA

²Center for Complex Biological Systems, University of California, Irvine, Irvine, USA

³Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, USA

⁴Department of Biomedical Engineering, Johns Hopkins University, Baltimore, USA

⁵McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University, Baltimore, USA

⁶Program in Computational Biology and Bioinformatics, Yale University, New Haven, USA

⁷Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, USA

⁸Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain

⁹Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, USA

¹⁰Department of Genetics, Stanford University School of Medicine, Palo Alto, USA

¹¹The Jackson Laboratory, The Jackson Laboratory, Bar Harbor, USA

¹²Cellular and Molecular Biology Program, University of Michigan, Ann Arbor, USA

¹³Department of Radiation Oncology, University of Michigan, Ann Arbor, USA

¹⁴Department of Biostatistics, University of Michigan, Ann Arbor, USA

¹⁵Center for RNA Biomedicine and Rogel Cancer Center, University of Michigan, Ann Arbor, USA

¹⁶Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, USA

¹⁷Software Engineering, insitro, South San Francisco, USA

¹⁸Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, USA

¹⁹Department of Neurological Sciences, Rush University Medical Center, Chicago, USA

²⁰Departments of Radiation Oncology and Environmental Health Sciences, University of Michigan, Ann Arbor, USA

²¹Section on Biomedical Informatics and Data Science, Yale University, New Haven, USA

²²Department of Statistics and Data Science, Yale University, New Haven, USA

²³Department of Computer Science, Yale University, New Haven, USA

²⁴Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, USA

²⁵Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain

The majority of mammalian genes encode multiple transcript isoforms that result from differential promoter use, changes in exonic splicing, and alternative 3' end choice. Detecting and quantifying transcript isoforms across tissues, cell types, and species has been extremely challenging because transcripts are much longer than the short reads normally used for RNA-seq. By contrast, long-read RNA-seq (LR-RNA-seq) gives the complete structure of most transcripts. We sequenced 264 LR-RNA-seq PacBio libraries totaling over 1 billion circular consensus reads (CCS) for 81 unique human and mouse samples. We detect at least one full-length transcript from 87.7% of annotated human protein coding genes and a total of 200,000 full-length transcripts, 40% of which have novel exon junction chains.

To capture and compute on the three sources of transcript structure diversity, we introduce a gene and transcript annotation framework that uses triplets representing the transcript start site, exon junction chain, and transcript end site of each transcript. Using triplets in a simplex representation demonstrates how promoter selection, splice pattern, and 3' processing are deployed across human tissues, with nearly half of multi-transcript protein coding genes showing a clear bias toward one of the three diversity mechanisms. Evaluated across samples, the predominantly expressed transcript changes for 74% of protein coding genes. In evolution, the human and mouse transcriptomes are globally similar in types of transcript structure diversity, yet among individual orthologous gene pairs, more than half (57.8%) show substantial differences in mechanism of di-

versification in matching tissues. This initial large-scale survey of human and mouse long-read transcriptomes provides a foundation for further analyses of alternative transcript usage, and is complemented by short-read and microRNA data on the same samples and by epigenome data elsewhere in the ENCODE4 collection.

INTRODUCTION

Most mammalian genes produce multiple distinct transcript isoforms¹. This transcript structure diversity is governed by promoter selection, splicing, and polyA site selection, which respectively dictate the transcript start site (TSS), exon junction chain (the unique series of exon-exon junctions used in a transcript), and transcript end site (TES, and the resulting 3' UTR) used in the final transcript. Each of these processes is highly regulated and is subject to a different set of evolutionary pressures^{2–5}. In protein coding genes, missplicing can lead to nonfunctional transcripts by disrupting canonical reading frames or introducing premature stop codons that predispose the transcript to nonsense mediated decay (NMD). Conversely, the cellular machinery involved in promoter or polyA site selection for protein coding genes is only constrained by the need to include start and stop codons for the correct open reading frame (ORF) in the final mRNA

product.

Transcript structure diversity poses challenges for both basic and preclinical biology. As computational gene prediction and manual curation efforts have identified ever more transcripts for many genes^{6,7}, a common assumption in genomics and medical genetics is that we only need to consider one or at most a handful of representative transcripts per gene such as those from the MANE (Matched Annotation from NCBI and EMBL-EBI) project⁸. MANE transcripts are chosen with respect to their expression levels in biologically-relevant samples and sequence conservation of the coding regions, and are perfectly matched between NCBI and ENSEMBL with explicit attention to the 5' and 3' ends. This decision to focus on one transcript per gene was driven in part by the difficulties in transcript assembly using ESTs and short-read RNA-seq, which is the assay used for most bulk and single-cell RNA-seq experiments^{9,10}. The advent of long-read platforms heralded the promise of full-length transcript sequencing to identify expressed transcript isoforms, thus potentially bypassing the error-prone transcript assembly step^{11,12}. However, as long-read RNA-seq (LR-RNA-seq) produces more novel candidate transcripts, there is a need to find organizational principles that will allow us to cope with the diversity of transcripts observed at some gene loci in catalogs such as GENCODE⁷, while at the same time distinguishing the genes that do not seem to undergo any alternative splicing.

Short-read RNA-seq has been the core assay for measuring gene expression in the second and third phases of the ENCODE project for all RNA biotypes, regardless of their lengths, in both human and mouse samples^{13–15}. Short-read RNA-seq has also been used by many groups to comprehensively characterize TSS usage¹⁶, splicing¹⁷, and TES usage¹⁸, but the challenges of transcript assembly given the combinatorial nature of the problem have precluded a definitive assessment of the transcripts present. In addition to continuing Illumina-based short-read sequencing of mRNA and microRNA, the fourth phase of ENCODE (ENCODE4) adds matching LR-RNA-seq using the Pacific Biosciences Sequel 1 and 2 platforms in a set of human and mouse primary tissues and cell lines in order to identify and quantify known and novel transcript isoforms expressed across a diverse set of samples. We report the resulting ENCODE4 human and mouse transcriptome datasets. We implement a novel triplet scheme that captures essential differences in 5' end choice, splicing, and 3' usage, which allows us to categorize genes based on features driving their transcript structure diversity using a new software package called Cerberus. We introduce the gene structure simplex as an intuitive coordinate system for comparing transcript usage between genes and across samples. We then compare transcript usage between orthologous genes in human and mouse and identify substantial differences in transcript diversity for over half the genes.

Results

The ENCODE4 RNA dataset. This LR-RNA-seq study profiled 81 tissues or cell lines by using the PacBio sequenc-

ing platform on 264 human and mouse libraries that include replicate samples and multiple human tissue donors (Tables S1-2). Without consideration for the seven postnatal time-points in mouse, they represent 49 unique tissues or cell types across human and mouse (Fig. 1a, Fig. S1). In addition, we sequenced matching human short-read RNA-seq (Fig. S1c) and microRNA-seq (Fig. S2, Supplementary results) for most samples as well as for an additional 37 that were sequenced with short-read RNA-seq only. We detect the vast majority of polyA genes (those with biotype protein coding, pseudogene, or lncRNA) whether we restrict the analysis to short-read samples that have matching data in the LR-RNA-seq dataset (93.9% of GENCODE v40 polyA genes and 90.6% of protein coding genes) or if we use all of the short-read samples (Fig. 1b, Fig. S3a). 31.1% of all expressed genes are detected in most (>90%) of the samples, and 34.0% are detected more specifically (<10% of samples) (Fig. 1c, Fig. S3b).

For each LR-RNA-seq dataset, we first mapped the reads using Minimap2¹⁹ and corrected non-canonical splice junctions and small indels using TranscriptClean²⁰, after which we ran TALON²¹ and LAPA²² to identify each transcript by its exon junction chain and assign each transcript a supported 5' and 3' end. Finally, to catalog transcript features and summarize transcript structure diversity in our datasets, we ran Cerberus, which is described below. It is important to emphasize that this pipeline (Fig. 1d) does not attempt to assemble reads, so that every reported known transcript is observed from 5' to 3' end in at least one read. We further required support from multiple reads for defining valid ends. Overall this is a conservative pipeline that was designed to detect and quantify robust novel and known transcripts (Materials and Methods).

Our LR-RNA-seq reads are oligo-dT primed and we therefore expect to see high detection of transcripts from polyA genes, which we define as belonging to the protein coding, lncRNA, or pseudogene GENCODE-annotated biotypes, across our datasets. Consistent with this expectation, we detect 75.9% of annotated GENCODE v40 polyA genes and 93.7% of protein coding genes at ≥ 1 TPM in at least one library in our human dataset (Fig. 1e). The overwhelming majority of undetected polyA genes are pseudogenes and lncRNAs, which are likely to be either lowly expressed or completely unexpressed in the tissues assayed. As expected, GO analysis of the undetected protein coding genes yielded biological processes such as smell and taste-related sensory processes that represent genes specifically expressed in tissues that we did not assay (Fig. S3c). We find that many genes are either expressed in a sample-specific manner (27.8% in <10% of samples) or are ubiquitously expressed across many samples (28.2% in >90% of samples), consistent with the short-read samples (Fig. 1f).

Transcriptionally active regions that are absent from GENCODE are candidates for novel genes. Applying conservative thresholds that included a requirement for one or more reproducible splice junctions (Supplementary methods), we found 214 novel candidate genes with at least one spliced

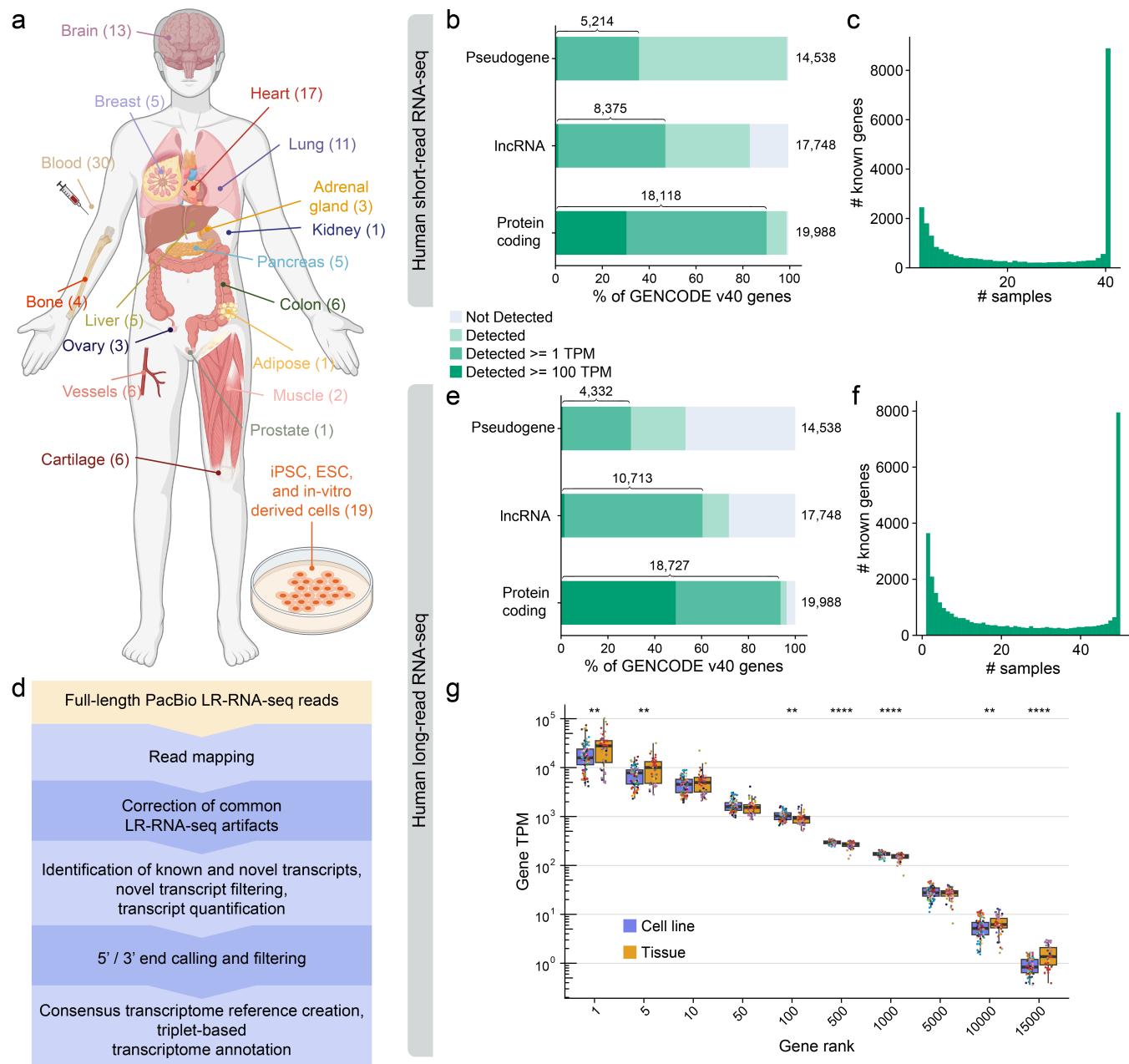


Figure 1. Overview of the ENCODE4 RNA datasets. **a**, Overview of the sampled tissues and number of libraries from each tissue in the ENCODE human LR-RNA-seq dataset. **b**, Percentage of GENCODE v40 polyA genes by gene biotype detected in at least one ENCODE short-read RNA-seq library from samples that match the LR-RNA-seq at > 0 TPM, >= 1 TPM, and >= 100 TPM. **c**, Number of samples in which each GENCODE v40 gene is detected >= 1 TPM in the ENCODE short-read RNA-seq dataset from samples that match the LR-RNA-seq. **d**, Data processing pipeline for the LR-RNA-seq data. **e**, Percentage of GENCODE v40 polyA genes by gene biotype detected in at least one ENCODE human LR-RNA-seq library at > 0 TPM, >= 1 TPM, and >= 100 TPM. **f**, Number of samples in which each GENCODE v40 gene is detected >= 1 TPM in the ENCODE human LR-RNA-seq dataset. **g**, Boxplot of TPM of polyA genes at the indicated rank in each human LR-RNA-seq library. Not significant (no stars) P > 0.05; *P <= 0.05, **P <= 0.01, ***P <= 0.001, ****P <= 0.0001; Wilcoxon rank-sum test.

transcript isoform expressed >= 1 TPM in human and 96 in mouse at our existing sequencing depth. Applying the same criteria to annotated polyA genes, we find 20,716 and 18,971 genes for human and mouse respectively, meaning that plausible novel genes constitute less than 1.0% in human and 0.5% in mouse. We subsequently focus analysis on transcripts from known polyA genes.

We then examined the distribution of gene expression values across our human LR-RNA-seq dataset to characterize the abundance of genes and to assess whether we would

be able to measure differences in transcript abundance at our current sequencing depths. For each library, we ranked each gene by TPM and found that the most highly expressed genes have higher TPMs in primary tissue-derived libraries than cell line-derived libraries (Fig. 1g, Supplementary methods). In particular, the tissue-derived liver libraries have the most highly expressed genes at ranks 1, 5, and 10, which include *ALB* and *FTL*, as expected. We also observed that the top 1,000 genes expressed in all but one liver library are expressed >= 100 TPM and that the top 5,000 are expressed >=

10 TPM. We can therefore confidently measure major transcript expression usage with a conservative threshold of 10 TPM for at least a third of expressed genes in each sample.

From genes expressed ≥ 10 TPM, we are able to capture over half (54.0%) of MANE transcripts that are 9-12 kb long (Fig. S3d). Coupled with our read length profiles, we estimate that we can reliably sequence the 99.7% of annotated GENCODE v40 polyA transcripts that are < 12 kb long from end-to-end if they are highly expressed (Fig. S3d-g). In mouse, we observe similar read length profiles, sample separation, and gene detection patterns (Fig. S3h-j), including detection of 84.9% of annotated GENCODE vM25 protein coding genes at ≥ 1 TPM (Fig. S3h). In summary, we are able to detect most of human and mouse protein coding genes in our ENCODE LR-RNA-seq datasets at similar rates to short-read RNA-seq, and our long reads are long enough to capture the vast majority of annotated polyA transcripts.

Different sources of transcript structure diversity. We compared the transcript start sites (TSSs), exon junction chains (ECs), and transcript end sites (TESs) observed in the human LR-RNA-seq data with other prior assays of these features and with established catalogs of these features. Cerberus is designed to identify unique TSSs, ECs, and TESs from a wide variety of inputs that include LR-RNA-seq data, reference atlases, and external transcriptional assays such as CAGE, PAS-seq, and the GTEx LR-RNA-seq dataset^{18,23} (Supplementary methods). Cerberus numbers each TSS, EC, and TES (triplet features) based on the annotation status of the transcript that it came from (e.g. the most confidently annotated will be numbered first) as well as the order in which each source was provided (Supplementary methods). Cerberus outputs genomic regions for each unique TSS and TES and a list of coordinates for each unique EC. In all cases, the gene of origin is also annotated (Fig. S4a). Using the integrated series of cataloged triplet features, Cerberus assigns a TSS, EC, and TES to each unique transcript model to create a transcript identifier of the form Gene[X,Y,Z], which we call the transcript triplet (Fig. S4b, Fig. 2a). This strategy distinguishes the structure of two different transcripts from the same gene solely on the basis of their transcript triplets. Additionally, it gives us the ability to sum up the expression of TSSs, ECs, and TESs across the transcripts they come from to enable quantification of promoter usage, EC usage, and polyA site usage respectively.

We applied Cerberus to the ENCODE human LR-RNA-seq data and to annotations from GENCODE v40 and v29 to obtain transcript triplets for the transcripts present in each transcriptome. Cerberus labels each triplet feature as known if it is detected in a reference set (here defined as transcripts derived from GENCODE) or novel if not. Additionally, using the information from the GENCODE reference transcriptomes, we assign the triplet [1,1,1] to the MANE transcript isoform for the gene, if it has one.

Altogether, we detected 206,806 transcripts expressed ≥ 1 TPM from polyA genes, 76,469 of which have exon junction chains unannotated in GENCODE v29 or v40. From these transcripts, we first sought to characterize the observed

triplet features (expressed ≥ 1 TPM in at least one library from polyA genes) in our dataset (Fig. 2b-d, Fig. S5a-f). We found that 18.0% of TSSs, 37.3% of ECs, and 22.1% of TESs are novel compared to both GENCODE v29 and v40 (Fig. 2b-d). We furthermore determined whether any novel triplet features were supported by sources outside of the GENCODE reference. We used CAGE and RAMPAGE data to support TSSs, GTEx transcripts to support ECs, and PAS-seq and the PolyA Atlas regions to support the TESs (Supplementary methods). Of the novel triplet features, 42.8% of TSSs, 17.9% of ECs, and 79.0% of TESs were supported by at least one external dataset (Fig. 2b-d, Fig. S5a-c). While the intermediate transcriptome (in general transfer format; GTF) for our LR-RNA-seq dataset from LAPA has single-base transcript ends, the majority of our Cerberus TSSs and TESs derived from the LR-RNA-seq data are 101 bp in length and 99.8% are shorter than 500 bp, which is consistent with how Cerberus extends TSSs and TESs derived from GTFs by n bp (here $n=50$) on either side (Fig. S4a, Fig. S5d-e, Supplementary methods).

We further annotated the novelty of ECs compared to GENCODE using the nomenclature from SQANTI²⁴. For detected ECs from polyA genes, we find that the majority (62.7%) of ECs were already annotated in either GENCODE v29 or v40. Novel ECs are primarily annotated as either NIC (16.1%; novel in catalog, defined as having a novel combination of known splice sites) or NNC (11.6%; novel not in catalog, defined as having at least one novel splice site) (Fig. S5f). Given the high external support for our triplet features, we were also able to predict CAGE and RAMPAGE support for our long-read derived TSSs both in and across cell types using logistic regression (Fig. S6, Supplementary results). In aggregate, a majority of our triplet features observed in our LR-RNA-seq data show were in prior annotations or have external support from additional assays.

We then examined the number of observed triplet features per gene. We find that most protein coding genes (89.8%) express more than one triplet feature across our dataset (Fig. 2e-g). By contrast, only 33.7% of lncRNAs and 14.4% of pseudogenes express more than 1 transcript and therefore triplet feature per gene. These biotypes exhibit far less transcript structure diversity as compared to protein coding genes. Overall, we find that our observed triplet features are individually well-supported by external annotations and assays. We also show that protein coding genes are far more likely to have more than one triplet feature than lncRNAs and pseudogenes.

The ENCODE4 LR-RNA-seq transcriptome. Following our characterization of individual triplet features, we moved on to examining our full-length transcripts. We first note that most observed transcripts with known ECs belong to the protein coding biotype (Fig. 2h). In contrast to the gene-level analysis, transcripts with known ECs are expressed in a more sample-specific manner, with 49.0% expressed in $< 10\%$ of samples and 4.4% expressed in $> 90\%$ of samples (Fig. S7a, Fig. 1f). Of the remaining protein coding transcripts with novel ECs, 53.0% are predicted to have com-

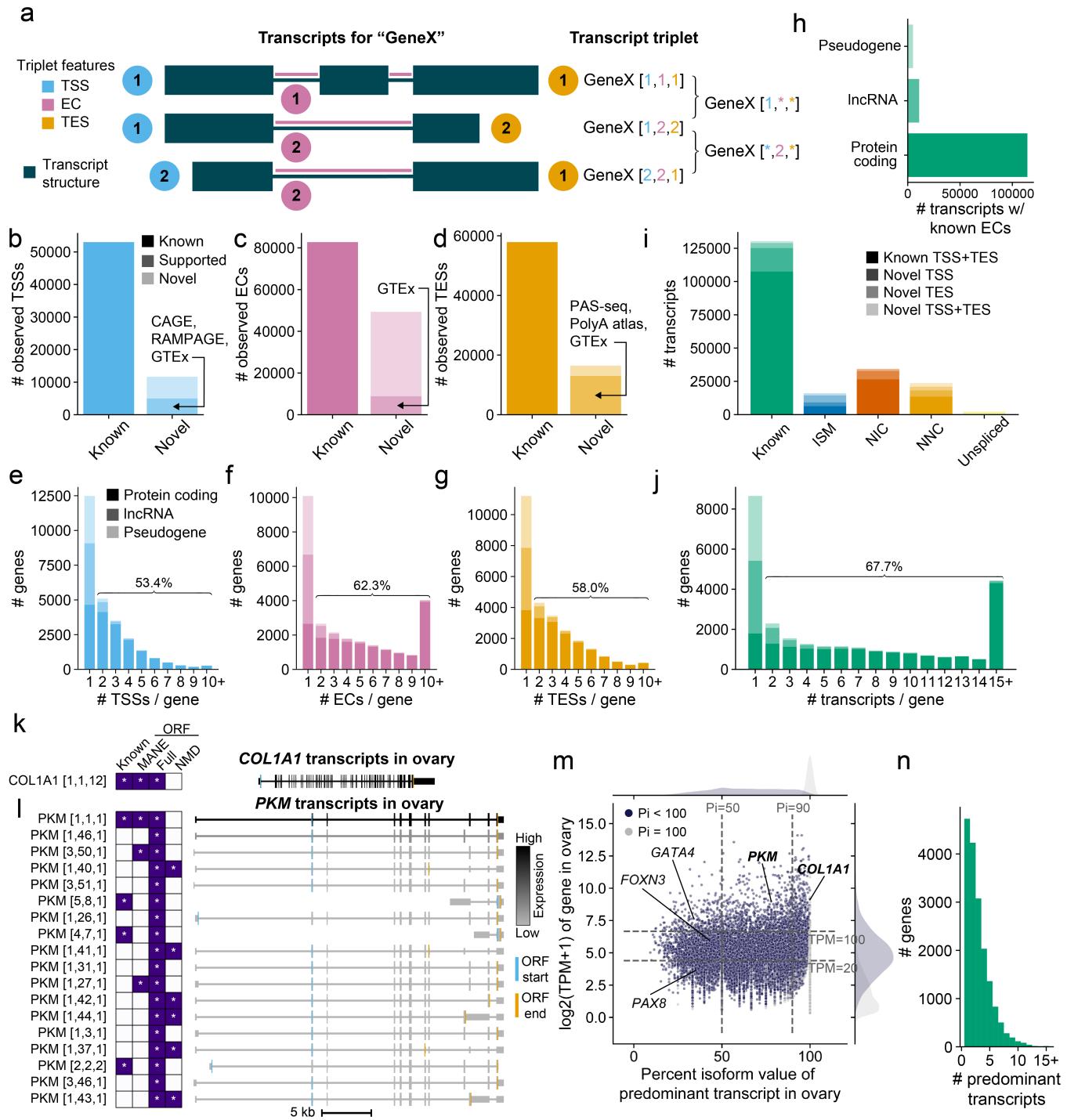


Figure 2. Triplet annotation of transcript structure maps diversity within and across samples. **a**, Representation of structure and transcript triplet naming convention for 3 different transcripts from the same gene based on the transcript start site (TSS), exon junction chain (EC), and transcript end site (TES) used. **b-d**, Triplet features detected ≥ 1 TPM in human ENCODE LR-RNA-seq from GENCODE v40 polyA genes broken out by novelty and support. Known features are annotated in GENCODE v29 or v40. Novel supported features are supported by **b**, CAGE or RAMPAGE **c**, GTEx, **d**, PAS-seq or the PolyA Atlas. **e-g**, Triplet features detected ≥ 1 TPM in human ENCODE LR-RNA-seq per GENCODE v40 polyA gene split by gene biotype for **e**, TSSs, **f**, ECs, **g**, TESs. **h**, Number of transcripts from GENCODE v40 polyA genes detected ≥ 1 TPM from human ENCODE LR-RNA-seq that have a known EC split by gene biotype. **i**, Novelty characterization of triplet features in each transcript detected ≥ 1 TPM in the human ENCODE LR-RNA-seq. **j**, Number of transcripts detected ≥ 1 TPM in human ENCODE LR-RNA-seq per GENCODE v40 polyA gene split by gene biotype. **k**, *COL1A1* (gene expressed at 548 TPM) transcripts expressed ≥ 1 TPM in the ovary sample from human ENCODE LR-RNA-seq. **l**, *PKM* (gene expressed at 506 TPM) transcripts expressed ≥ 1 TPM in the ovary sample from human ENCODE LR-RNA-seq colored by expression level (TPM). **m**, Expression level of gene (TPM) versus the percent isoform (p_i) value of the predominant transcript for each gene expressed ≥ 1 TPM from human ENCODE LR-RNA-seq in the ovary sample. Points are colored by whether or not $p_i = 100$. **n**, Number of unique predominant transcripts detected ≥ 1 TPM across samples per gene.

plete ORFs which are not subject to nonsense mediated decay (Supplementary methods). Examining detected transcripts in our dataset based on the novelty of each constituent triplet feature, we find that 52.0% of observed transcripts have each of their triplet features annotated, and that more transcripts contain novel TSSs than novel TESs (Fig. 2i). Consistent with our observation that protein coding genes generally have more than one triplet feature per gene compared to the other polyA biotypes, we find that most protein coding genes also have more than one transcript per gene (Fig. 2j, e-g). We investigated the extent that lower expression levels of lncRNAs contribute to their overall lower transcript diversity compared to protein coding genes. We found that the 137 lncRNAs expressed >100 TPM in one or more samples have the same median number of expressed transcripts per gene as the 8,436 protein coding genes at the same expression level (median = 7) (Fig. S7b, Supplementary methods). Therefore the lower reported overall diversity of lncRNAs is due to a combination of their lower expression levels and our sequencing depth.

We compared the number of TSSs and TESs that are detected per EC across the observed transcripts from GENCODE v40 versus in our observed transcripts. We found that in GENCODE v40, each multixonic EC has a maximum of 3 TSSs or TESs across the polyA transcripts and that the overwhelming majority of ECs are only annotated with 1 TSS and TES (99.7% and 99.4% respectively) (Fig. S7c-d). In contrast, our strategy of transcriptome annotation yields a substantial increase in the number of distinct TSSs and TESs observed per EC, which more accurately reflects the biology of the coordination of promoter choice, polyA site selection, and splicing across the diverse samples in our dataset (Fig. S7e-f). The effect of this increase in annotated TSSs and TESs is also apparent when analyzing our transcripts using traditional alternative splicing event detection methods, which are not written to consider the more subtle differences in transcript structure at the 5' and 3' end (Fig. S8, Supplementary results).

Predominant transcript structure differs across tissues and cell types. Different multixonic genes with similar expression levels within the same sample can exhibit vastly different levels of transcript structure diversity. For instance, the genes *COL1A1* and *PKM* have a high number of exons (60 and 47 exons, respectively across our entire human dataset) and are highly expressed in ovary (548 and 506 TPM respectively). Yet, we detect only one 6.9 kb long transcript for *COL1A1* (Fig. 2k) whereas we detect 18 transcript isoforms that vary on the basis of their TSSs, ECs, and TESs for *PKM* (Fig. 2l).

We then asked what fraction of overall gene expression is accounted for by the predominant transcript, which is the most highly expressed transcript for a gene in a given sample. Comparing the TPM of genes expressed in ovary to the percentage of reads from a gene that come from that transcript (π - percent isoform)²⁵ of the predominant transcript, we find that 19.5% of protein coding genes expressed >100 TPM have a predominant transcript that accounts for less than 50% of the reads, and therefore are highly expressed with

high transcript structure diversity. Conversely, 26.8% of protein coding genes are expressed >100 TPM and have a predominant transcript that accounts for more than 90% of the expression of the gene (Fig. 2m). Globally, we generated a catalog of predominant transcripts for each sample. The median number of predominant transcripts per protein coding gene across samples was 2, and that 73.0% of protein coding genes have more than one predominant transcript across the samples surveyed (Fig. 2n). Thus, the majority of human protein coding genes use a different predominant transcript in at least one condition represented in our sample collection.

Quantifying transcript structure diversity across samples using gene triplets and the gene structure simplex. We developed a framework to systematically characterize and quantify the diversity between the detected transcripts from each gene by computing a summary gene triplet, which is related to but distinct from transcript triplets. For each set of transcripts from a given gene, we count the number of unique TSSs, ECs, and TESs (Fig. 3a, Fig. S9). As the number of exon junction chains is naturally linked to the number of alternative TSSs or TESs (for instance, a new TSS with a different splice donor will lead to a novel EC regardless of similarities in downstream splicing), we calculate the splicing ratio as $\frac{2 \times N_{EC}}{N_{TSS} + N_{TES}}$ to more fairly assess the contribution of ECs to transcript diversity in each gene (Fig. 3a, Fig. S9). We then compute the proportion of transcript diversity that arises from each source of variation: alternative TSS usage, alternative TES usage, or internal splicing (Fig. 3a). Representing these numbers as proportions allows us to plot them as coordinates in a two-dimensional gene structure simplex (Fig. 3b, Fig. S9). This enables us to visualize how transcripts from a gene typically differ from one another and categorize genes based on their primary driver of transcript structure diversity. Genes with a high proportion of transcripts characterized by alternative TSS usage (>0.5) will fall into the TSS-high sector of the simplex, those with a high proportion of transcripts characterized by alternative TES usage (>0.5) will fall into the TES-high sector of the simplex, and those with a high proportion of transcripts characterized by internal splicing (>0.5) will fall into the splicing-high portion of the simplex. Genes with more than one transcript that do not display a strong preference for one mode over the other lie in the mixed sector, and genes with just one transcript are in the center of the simplex, henceforth the simple sector (Fig. 3a-b, Fig. S9, Supplementary methods).

We first used the gene structure simplex to compare different transcriptomes. We computed gene triplets for protein coding genes for the following transcriptomes: GENCODE v40 transcripts from genes we detect in our LR-RNA-seq dataset; observed transcripts in our LR-RNA-seq dataset (observed); and the union of detected major transcripts (observed major), which we define as the set of most highly expressed transcripts per gene in a sample that are cumulatively responsible for over 90% of that gene's expression in any of our LR-RNA-seq samples (Fig. 3c-j, Supplementary methods). The observed and observed major gene triplets describe the diversity of transcription in each gene across all samples in the

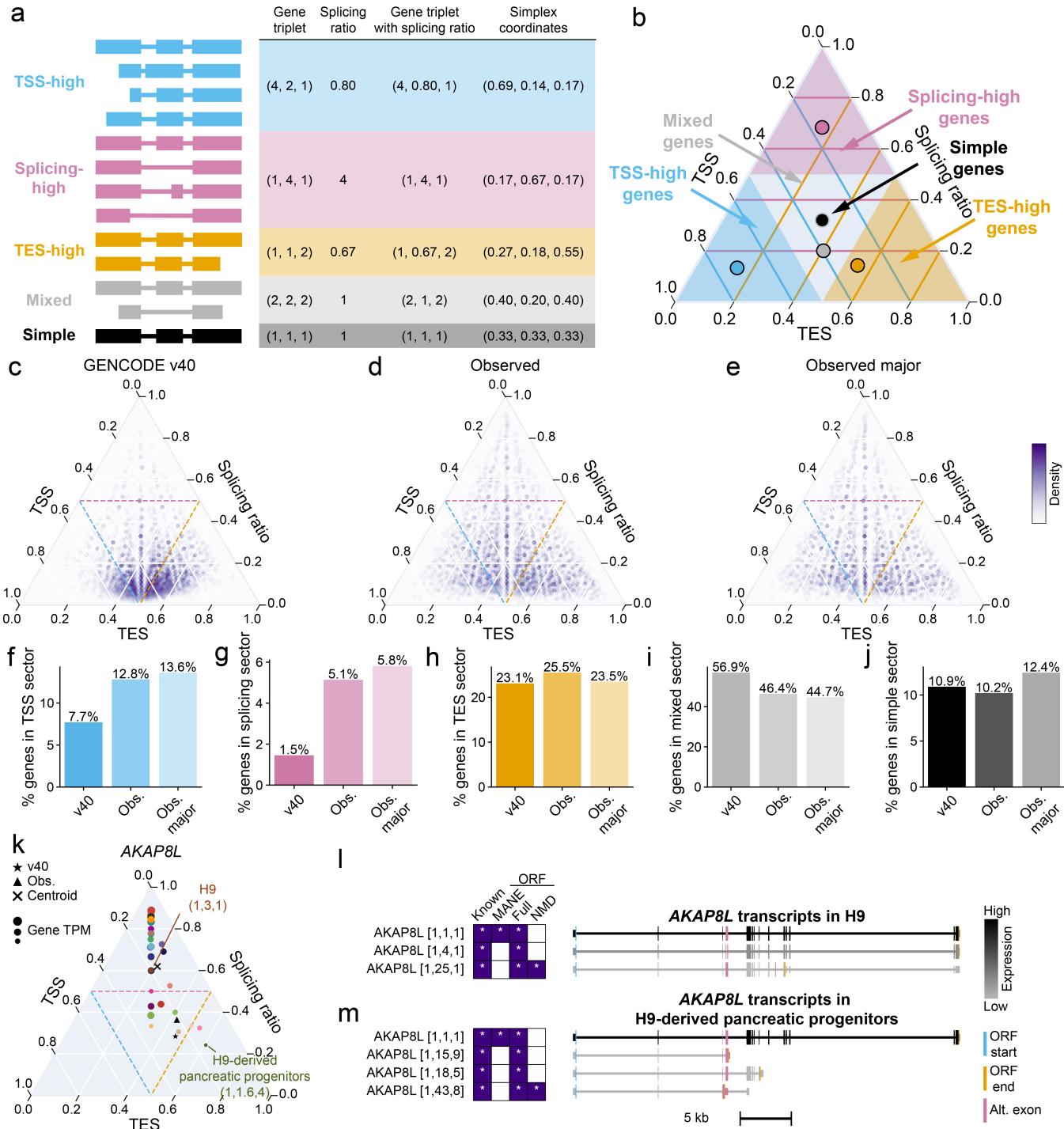


Figure 3. The gene structure simplex represents distinct modes of transcript structure diversity across genes and samples. **a**, Transcripts for 5 model genes; 1 of each sector (TSS-high, splicing-high, TES-high, mixed, and simple). Table shows the gene triplet, splicing ratio gene triplet, and simplex coordinates that correspond to each toy gene. **b**, Layout of the gene structure simplex with the genes from **a**, plotted based on their simplex coordinates. Proportion of TSS usage is the blue axis (left), proportion of TES is the orange axis (bottom), and proportion of splicing ratio is the pink axis (right). Regions of the simplex are colored and labeled based on their sector category (TSS-high, splicing-high, TES-high). Gene triplets that land in each sector are assigned the concordant sector category. **c-e**, Gene structure simplices for the transcripts from protein coding genes that are **c**, annotated in GENCODE v40 where the parent gene is also detected in our human LR-RNA-seq dataset, **d**, the observed set of transcripts, those detected ≥ 1 TPM in the human ENCODE LR-RNA-seq dataset, **e**, the observed major set of transcripts, the union of major transcripts from each sample detected ≥ 1 TPM in the human ENCODE LR-RNA-seq dataset. **f-j**, Proportion of genes from the GENCODE v40, observed, and observed major sets that fall into the **f**, TSS-high sector, **g**, splicing-high sector, **h**, TES-high sector, **i**, mixed sector, **j**, simple sector. **k**, Gene structure simplex for AKAP8L. Gene triplets with splicing ratio for H9 and H9-derived pancreatic progenitors labeled. Simplex coordinates for the GENCODE v40, observed set, and centroid of the samples also shown for AKAP8L. **l-m**, Transcripts of AKAP8L expressed ≥ 1 TPM in **l**, H9 **m**, H9-derived pancreatic progenitors colored by expression level in TPM. Alternative exons that differ between transcripts are colored pink.

dataset. Unsurprisingly, GENCODE genes show less density in the TSS or TES sectors of the simplex, largely because the main focus of GENCODE is to annotate unique ECs rather than 5' or 3' ends. This causes a concomitant drop in diversity in these sectors in GENCODE compared to the observed and observed major transcripts (Fig. 3f, h). Interestingly, there is also a distinct enrichment of genes that occupy the splicing-high portion of the simplex in our observed set compared to GENCODE (Fig. 3g). When considering the observed major transcripts, we see an increase in the percentage of genes in the TSS and splicing-high sectors over the set of all transcripts detected in our entire LR-RNA-seq dataset, but a decrease in the TES-high sector (Fig. 3f-h). Overall, we compared gene triplets for transcripts as annotated by GENCODE and observed in our LR-RNA-seq dataset and found higher proportions of genes with high TSS and splicing diversity as compared to GENCODE.

Calculating sample-level gene triplets identifies genes that show distinct transcript structure diversity across samples. The observed gene triplets represent the aggregate repertoire of triplet features for each gene globally across our entire LR-RNA-seq dataset. However, the overall transcript structure diversity of a gene does not necessarily reflect the transcript structure diversity of a gene within a given sample. Therefore, we computed gene triplets for each sample in our dataset using all detected transcripts in each sample (sample-level gene triplets) or just the major transcripts in each sample (sample-level major gene triplets). These gene triplets can also be visualized on the gene structure simplex where each point represents the gene triplet associated with a different sample (Fig. 3k).

In order to find genes that display heterogeneous transcript structure diversity across unique biological contexts, we computed the average coordinate (centroid) for each gene from the sample-level gene triplets and calculated the distance between it and each sample-level gene triplet (Supplementary methods). 2,892 unique genes had a distance z-score >3 in at least one sample and therefore demonstrate dissimilar transcript structure diversity from the average. One such example is *AKAP8L* in the H9-derived pancreatic progenitors (z-score: 5.23). *AKAP8L* can bind both DNA and RNA in the nucleus and has been shown to have functional differences on the protein level resulting from alternate transcript choice^{26,27}. In our data, transcripts of this gene generally differ in terms of the EC or TES choice, but this behavior differs from sample to sample (Fig. 3k). For example, transcripts of *AKAP8L* differ only in their ECs in H9 embryonic stem cells, whereas transcripts differ in their ECs and TESs in the H9-derived pancreatic progenitors (Fig. 3l-m).

We also compared our sample-level gene triplets to the observed gene triplets to understand how transcript structure diversity differs globally versus within samples (Fig. 3d, Fig. 2j, Fig. S10a-h). First, we simply counted the number of triplet features or transcripts per gene and found that while most genes have more than one triplet feature or transcript globally (Fig. 2e-g, j), on the sample level, most genes have far fewer triplet features and transcripts; with a particularly

pronounced difference for the TSS (Fig. S10a-d). We found that the distributions of triplet features overall and in each sample were significantly different from one another (two-sided KS test) (Fig. S10e-f, Supplementary methods).

To determine how transcript structure diversity for each gene changes from the global to sample level using the gene structure framework, we computed distances between the global observed gene triplets for non-simple genes to sample-level gene triplet centroids (Supplementary methods). We find that 3.2% of tested genes have a distance z-score >2 between their observed and sample-level centroid gene triplets. In support of our analysis on the individual triplet feature level, we find that 94.8% of genes from the TSS-high sector in the observed set do not share this sector with their sample-level centroid, indicating that genes with a large number of promoters typically use them in a sample-specific manner. *ACTA1*, a gene that encodes for an actin protein²⁸, is the gene with the highest distance between observed and sample-level centroid. Its observed gene triplet is (1,18,1) and therefore splicing-high. However, in most samples where *ACTA1* is expressed, it has only one transcript isoform (Fig. S10m). This drives the sample-level centroid behavior into the mixed sector (Fig. S10n). In contrast, in heart and muscle *ACTA1* expresses 18 and 15 transcripts respectively, which all differ on the basis of their ECs (Fig. S10m-o). This illustrates how the gene structure framework can be used to highlight differences between sample-specific and global transcript structure diversity, and also shows that individual genes are substantially different.

Sample-specific and global changes in predominant and major transcript isoform usage. Nevertheless, the transcript structure diversity pattern for the majority of genes is consistent across samples where they are expressed at substantial levels. Elastin (*ELN*), which is an important component of the extracellular matrix²⁹, is the gene with the greatest number of detected transcripts in our dataset (283 in total). We find that in most samples, distinct transcripts of *ELN* are characterized by different ECs (Fig. 4a). For example, in lung, *ELN* has 32 major transcripts with 21 different ECs, but in 31 of its major transcripts, uses only one TSS and two TESs (Fig. 4b). By contrast, the four transcripts from the transcription factor *CTCF* expressed in lung use three TSSs but only one TES (Fig. 4c-d).

While the observed gene triplets for a gene represent the overall transcript structure diversity, the observed major gene triplets capture diversity of the most highly expressed transcripts in each sample. We computed the distances between the observed and observed major simplex coordinates for protein coding genes. The transcription factor *E4F1*³⁰ has a high distance between the observed and observed major gene triplets, which corresponds to a change from the mixed to splicing-high sector (Fig. 4e-f). This sector change is driven by the use of fewer TSSs and TESs in major transcripts. Overall, 83.7% of protein coding genes retain their sectors between our observed and observed major triplets, while 4.8% genes in the mixed sector move to one of the three corners of the simplex (TSS, splicing, or TES-high) (Fig. 4g).

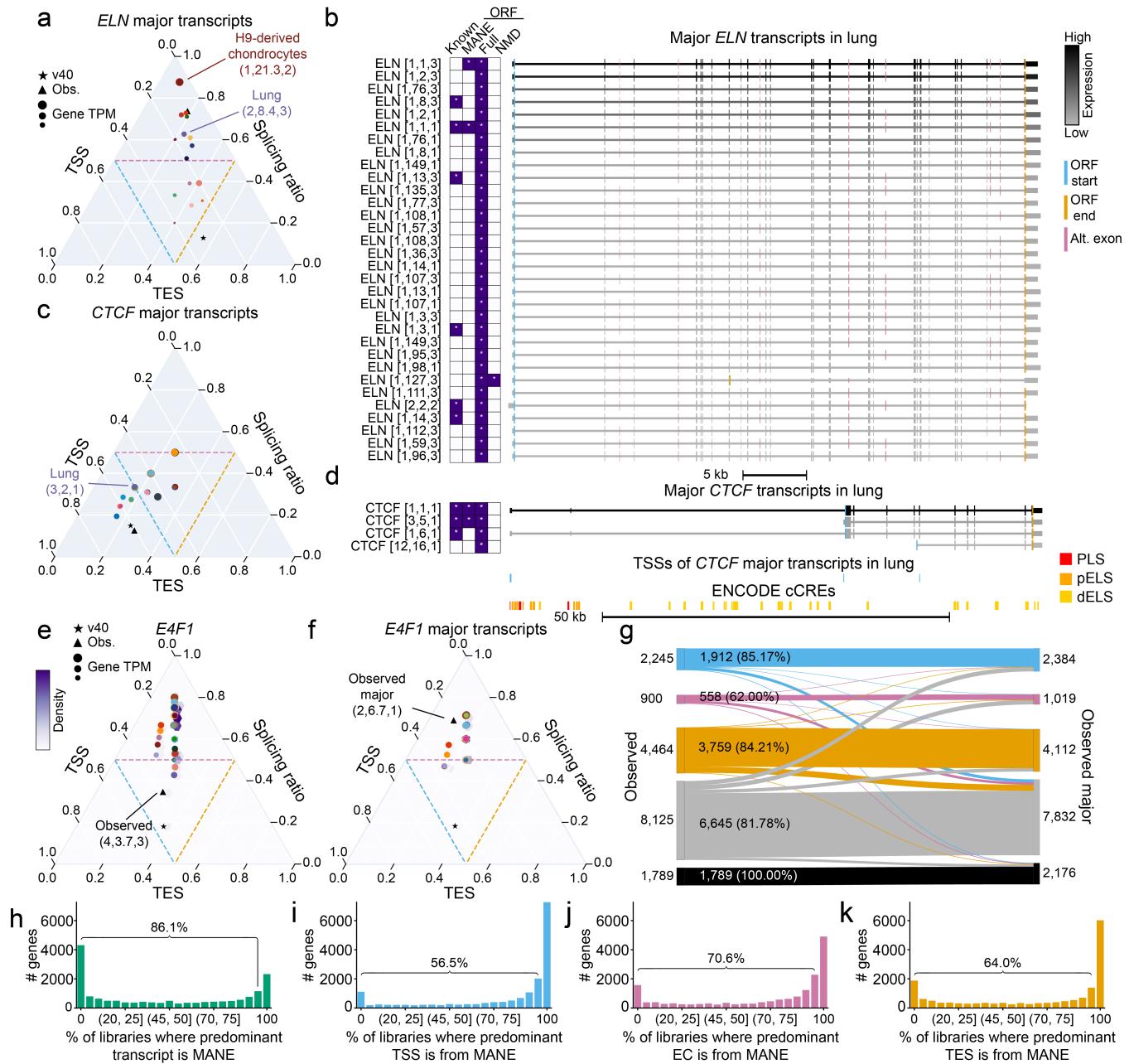


Figure 4. Sample-specific and global changes in predominant and major transcript isoform usage. **a**, Gene structure simplex for major transcripts of *ELN*. Gene triplets with splicing ratio for lung and H9-derived chondrocytes labeled. Simplex coordinates for the GENCODE v40 and observed major set are labeled. **b**, Major transcripts of *ELN* expressed ≥ 1 TPM in lung colored by expression level in TPM. Alternative exons that differ between transcripts are colored pink. **c**, Gene structure simplex for major transcripts of *CTCF*. Gene triplets with splicing ratio for lung labeled. Simplex coordinates for the GENCODE v40 and observed major set are labeled. **d**, From top to bottom: Major transcripts of *CTCF* expressed ≥ 1 TPM in lung, TSSs of *CTCF* major transcripts expressed ≥ 1 TPM in lung, ENCODE cCREs colored by type. **e**, Gene structure simplex for *E4F1*. Gene triplet with splicing ratio for observed *E4F1* transcripts labeled. Simplex coordinates for the GENCODE v40 and observed set also shown for *E4F1*. **f**, Gene structure simplex for major transcripts of *E4F1*. Gene triplet with splicing ratio for observed major *E4F1* transcripts labeled. Simplex coordinates for the GENCODE v40 and observed major set also shown for *E4F1*. **g**, Sector assignment change and conservation for protein coding genes in the human ENCODE LR-RNA-seq dataset between the observed set of gene triplets (left) and the observed major set of gene triplets (right). Percent of genes with the same sector between both sets labeled in the middle. **h-k**, Percentage of libraries where a gene with an annotated MANE transcript is expressed and the MANE **h**, transcript **i**, TSS **j**, EC **k**, TES is the predominant transcript or triplet feature.

Thus, the differences between the observed and observed major gene triplets in a subset of genes can be substantial.

One criterion for the identification of MANE transcripts is how highly expressed the transcript is compared to others⁸. Therefore, we assessed how frequently the MANE transcript was the predominant one in each of our LR-RNA-seq libraries. Limiting ourselves to only the genes that have annotated MANE transcripts in GENCODE v40, we found that 64.1% of genes have a non-MANE predominant transcript in at least 80% of the libraries where the gene is expressed (Fig. 4h). At the individual triplet feature level, 30.8% of TSSs, 40.9% of ECs, and 45.2% of TESs have a non-MANE predominant feature in at least 80% of libraries (Fig. 4i-k). Therefore, though the MANE transcript typically is the most highly expressed transcript in a library, most genes with MANE transcripts have some libraries where this is not the case. For non-MANE predominant transcripts, only 17.0% were predicted to have the same ORF as the MANE transcript. Furthermore, 62.1% of non-MANE predominant transcripts are predicted to encode for a full ORF that does not undergo NMD. These results indicate that in many cases, the alternative predominant transcript in a sample likely encodes for a distinct, functional protein. The genes where the MANE transcript and triplet features are frequently not the predominant one represent loci that would suffer more from restricting analyses to only a single transcript isoform.

For a subset of gene / library combinations where the MANE transcript or feature was not the predominant one, the MANE transcript or feature was still expressed, albeit at a lower level. For these gene / library combinations, we compared the expression of the predominant transcript to the MANE one (Fig. S11a-d). We found that for predominant transcripts or triplet features expressed <30 TPM, the MANE counterpart was expressed at a comparable level. By contrast, for the opposite situation, where the MANE transcript or triplet feature was the predominant one, we found that the secondary transcript was not expressed at a similar level (Fig. S11e-h). Overall, for most gene / library combinations, the MANE transcript or triplet feature is the predominant one (Fig. S11i-l).

Comparing transcript structure diversity between species. We ran Cerberus on the ENCODE4 mouse LR-RNA-seq dataset to calculate transcript and gene triplets to enable comparison of transcript structure diversity between the two species. Compared to human GENCODE v40, GENCODE vM25 genes are less enriched in the TSS, splicing, and TES-high sectors (Fig. 3f-j, Fig. 5a-e, Fig. S12a-c). For the mouse observed and observed major gene triplets, we see relatively similar percentages of genes in each sector (Fig. 5a-e, Fig. S12a-c). We found fewer predominant transcripts across samples per protein coding gene in mouse than in human, which is expected due to the overall lower number of tissues in our mouse data, with a median of 2 predominant transcripts per gene and 57.7% of protein coding genes with more than one (Fig. S12d). Furthermore, we observe that 54.5% of protein coding transcripts with novel ECs are predicted to encode for full ORFs without NMD. Thus, the two

transcriptomes have similar distributions of genes in our gene structure simplex.

In order to make gene-level comparison for orthologous genes in both species, we subset the human samples on those that are the most similar to the mouse samples and computed “mouse matched” observed, observed major, and sample-level gene triplets (Supplementary methods, Table S1-2). We computed the sample-level centroids for each gene in both species and computed the distance between each pair of 1:1 orthologs. Of the 13,536 orthologous genes, 4.3% have a distance z-score >2 between the species and therefore exhibit substantial changes in transcript structure diversity between the species. One of these is ADP-Ribosylation Factor 4 (*ARF4*), which is the most divergent member of the *ARF4* family³¹. Human *ARF4* sample-level gene triplets are nearly always splicing-high whereas mouse *Arf4* sample-level gene triplets are mainly TES-high (Fig. 5f-g). We examined the *ARF4* / *Arf4* transcripts expressed in matching embryonic stem cell samples (H1 in human and F121-9 in mouse) and found that, despite the homologous samples, all 3 of the expressed human transcripts use the same TSS and TES but differ in the ECs whereas all 3 expressed mouse transcripts use the same TSS and EC but differ at the TES (Fig. 5h-i). We find globally that when comparing the observed major gene triplets between human and mouse, only 42.2% of genes have the same sector in human and mouse (Fig. 5j). This result holds even when restricting ourselves to comparing human tissues with adult mouse samples or just a comparison between human and mouse embryonic stem cells (Fig. 5k). Thus, we find substantial differences in splicing diversity for orthologous genes between human and mouse.

Discussion

The ENCODE4 LR-RNA-seq dataset is the first large-scale, cross-species survey of transcript structure diversity using full-length cDNA sequencing on long-read platforms. We identify and quantify known and novel transcripts in a broad and diverse set of samples with uniformly processed data and annotations available at the ENCODE portal. A new framework was introduced for categorizing transcript structure diversity based on their exon junction chains and ends using gene and transcript triplets, which allowed us to use the gene structure simplex to visualize and compare gene triplets between samples and across species. The results showed a full range of transcript structure diversity across the transcriptome, based on promoter, internal splicing, and polyA site choice. As expected, the existing gene annotation catalogs such as GENCODE have successfully captured individual features such as TSSs and exons. However, GENCODE annotated full-length transcripts only represent a subset of the TSS, EC, and TES combinations that we observe using our conservative pipeline that requires full end-to-end support in a single read and support from multiple reads for defining ends. From the human LR-RNA-seq quantification, we found more than one predominant transcript across samples for 73.0% of genes, which is in contrast with prior reports^{32,33}. We also found that for a substantial number of

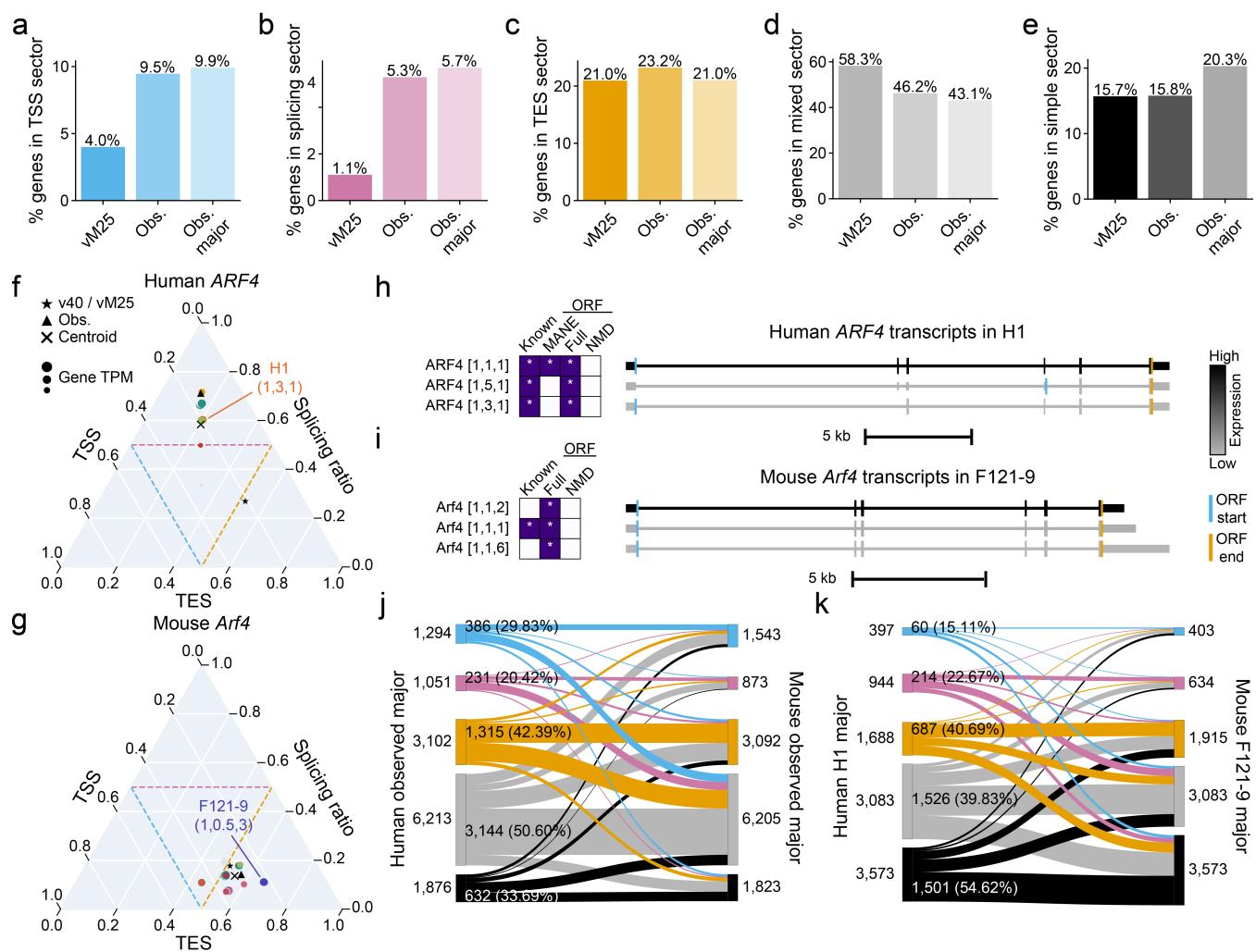


Figure 5. Conservation of gene triplets from human and mouse. **a-e**, Proportion of genes from the GENCODE vM25, observed, and observed major sets that fall into the **a**, TSS-high sector, **b**, splicing-high sector, **c**, TES-high sector, **d**, mixed sector, **e**, simple sector. **f**, Gene structure simplex for *ARF4* in human. Gene triplet with splicing ratio for *ARF4* transcripts in H1 sample labeled. Simplex coordinates for the GENCODE v40, sample-level centroid, and observed set also shown for *ARF4*. **g**, Gene structure simplex for *Arf4* in mouse. Gene triplet with splicing ratio for *Arf4* transcripts in F121-9 sample labeled. Simplex coordinates for the GENCODE v40, sample-level centroid, and observed set also shown for *Arf4*. **h**, Transcripts of *ARF4* expressed ≥ 1 TPM in human H1 sample colored by expression level in TPM. **i**, Transcripts of *Arf4* expressed ≥ 1 TPM in mouse F121-9 sample colored by expression level in TPM. **j**, Sector assignment change and conservation for orthologous protein coding genes between the observed major human set of gene triplets (left) and the observed major mouse set of gene triplets (right). Percent of genes with the same sector between both sets labeled in the middle. **k**, Sector assignment change and conservation for orthologous protein coding genes between the sample-level H1 major human set of gene triplets (left) and the sample-level F121-9 major mouse set of gene triplets (right). Percent of genes with the same sector between both sets labeled in the middle.

genes, transcript structure diversity and major transcript usage for the same gene differs between tissues, samples, and developmental timepoints. The majority of genes had at least one library where the MANE transcript is not the predominant transcript. This could confound analyses such as variant effect prediction in which it is common practice to consider only one transcript per gene. Finally, we found that transcript structure diversity behavior differs quite strikingly between human and mouse on a gene-by-gene basis. In matching samples, the dominant source of transcript structure diversity differed for more than half of orthologous protein coding genes.

Our data and framework provide a foundation for further analyses such as the functional impacts of alternative 5' and 3' ends, RNA modifications, RNA binding protein function, allele-specific expression, and transcript half-life. Together with the accompanying tissue and cell type annotations, this constitutes a transcript-level reference atlas that is structured

appropriately for integration of future single-cell long-read analysis. Exploration using the gene structure simplex analysis will yield additional genes showing sample specific variance compared to their average behavior when extended to new tissues, differentiation time courses, or disease samples. The triplet annotation scheme for transcripts, based on mechanistically distinct transcript features, organizes and simplifies high-level analysis of transcripts from the same gene. We find it to be a useful and commonsense improvement over arbitrary transcript IDs and we expect it to be widely applicable to transcriptomes of any organism that uses regulated alternative splicing, promoter choice, or 3' end selection.

Our annotations are consistent and extensive, yet they have several limitations. With the current sample preparation protocol and depth of sequencing, we reliably detect transcripts that are expressed above a minimum expression level of 1 TPM and are less than 10 kb long. While 99.3% of

GENCODE v40 polyA transcripts are less than 10 kb long, we are undoubtedly underrepresenting transcripts that are at the long end of the distribution, especially when they are expressed at low levels or in rare cell types within a tissue. RNA integrity differences between the human cell lines and mouse tissues, both of which produce very high quality RNA compared to human postmortem tissues, are expected to affect our results, because we have focused on full-length transcript sequencing rather than read assembly. Our imposition of minimum expression and inherent length limitations could also lead to lower sensitivity of splicing diversity in lncRNAs, which have thus far generated staggering transcript structure complexity when sequenced after enrichment capture³⁴. We also had lower detection of pseudogenes, and we hypothesize that the PacBio platform's accuracy and read length reduce the multimapping errors typical of short reads, especially for pseudogenes of highly expressed genes.

Within these boundaries outlined above, we were able to assess the sources and specifics of transcript structure diversity for major transcripts of most protein coding genes. Nearly all studies that have examined alternative splicing have emphasized transcript isoform multiplicity per gene. Also as expected, studies that have applied more permissive processing pipelines, used transcript assembly, or focused on nuclear RNA typically find evidence for far more RNA transcripts, especially at lower expression ranges^{24,35-38}. Assigning biological functions to new transcripts from our collection or any other contemporary study is a major challenge for the field. Unlike DNA replication's elaborate mechanisms to ensure fidelity, the three major processes of RNA biogenesis mapped here are understood to operate with less stringent fidelity, and though it has long been debated, we consider evidence for the existence of a new transcript isoform simply makes it a candidate of interest for a protein coding, precursor, or regulatory function.

The range of regulation used by different genes was illuminating. *COLIA1*, a complex gene in terms of number of exons, exhibited minimal transcript structure diversity in spite of high expression versus other genes of similar expression levels, such as *PKM* with its many transcripts resulting from all three mechanisms. This implies that transcript structure diversity is a property of the gene that has been optimized in evolution. This has major implications for evaluating the functions of regulatory factors such as *PAX6*, which has 81 transcripts in GENCODE v40, and 33 transcripts in our dataset. Conventional gene-level short-read RNA-seq profiling is likely obscuring important distinctions in transcript usage. While not every one of these transcripts leads to a difference in the protein product, changes at the 5' and 3' end are likely to alter the regulation of those transcripts. The incorporation of transcript usage as well as its regulation within the framework of gene regulatory networks, where appropriate, is a major challenge going forward.

Considering transcript structure diversity as a fundamental, tunable property of gene function, the mouse-human comparative results were the most surprising to us. In genomics and in the wider biology community we often use

orthology of mouse and human genes to predict and interpret gene function *in vivo*, including many uses of mice as mammalian models for both basic and preclinical purposes. The differences in transcript structure diversity that surfaced when we compared matching tissues from human and mouse suggests that this diversity is rapidly evolving on a per gene basis, even between primates and rodents. This is, however, consistent with prior observations of a large population of rapidly evolving candidate cis-regulatory elements³⁹. The results presented here provide a roadmap for evaluating the evolution of transcript structure diversity across species and impetus to focus on it, especially for genes with substantial differences that would affect interpretation of existing animal models and expectations for humanized gene-locus mouse models. It is hard to underestimate the need for better methods to test the functional significance of different transcript isoforms.

ACKNOWLEDGEMENTS

We thank the UCI GRTH for sequencing PacBio LR-RNA-seq libraries and the Caltech Jacobs Genetics and Genomics Laboratory for sequencing the Illumina mRNA RNA-seq libraries. We thank Gloria Sheynkman for guidance on protein prediction analyses.

A.M., B.J.W., L.R., D.B. were supported by UM1HG009443. B.J.W. was also supported by the Caltech Beckman Institute BIFGRC. D.B. was also supported by P30AG10161, P30AG72975, R01AG15819, R01AG17917, U01AG46152, U01AG61356. Z.W., M.B.G., R.G., and the members of the ENCODE DAC were supported by U24HG009446. M.L. was supported by UM1HG009382. M.B. was supported by R01HG012367 and U01HG009380. B.H. and the members of the ENCODE DCC were supported by U24HG009397.

AUTHOR CONTRIBUTIONS

B.W., G.B., E.R., H.Y.L., C.J.M., S.R., S.J., W.Z., K.C., A.K.W., and L.A.D. did the experimental work. F.R., D.W., M.H.C., E.R., N.R., D.T., S.R., J.S., S.U., J.J., I.Y., I.G., P.S., O.J., J.S.S., M.S.K., and B.C.H. developed the data processing pipelines. F.R., E.R., N.R., D.T., M.R., Y.J., B.B., S.M., A. McShane, K.B., and S.I.E. performed data analysis. F.R., B.W., E.R., N.R., M.R., Y.J., B.B., and A. Mortazavi wrote the manuscript. All other authors provided project input and edited the manuscript.

DATA AVAILABILITY

- Human LR-RNA-seq processed data / processing pipeline
- Human LR-RNA-seq datasets
- Mouse LR-RNA-seq processed data / processing pipeline
- Mouse LR-RNA-seq datasets
- Human short-read RNA-seq datasets
- Human microRNA-seq datasets

CODE AVAILABILITY

- Data processing and figure generation code
- Cerberus

SUPPLEMENTARY RESULTS

The ENCODE4 microRNA-seq dataset. We sequenced 254 human samples with microRNA-seq, which specifically captures mature (21-25 bp) microRNAs. We mapped our microRNA-seq data to pre-microRNA sequences in GENCODE v29. We detected 1,130 microRNAs at CPM ≥ 2 across the full dataset (Fig. S2a). Overall, microRNAs are more sample-specific than known genes in both short and long-read RNA-seq (Fig. S2a, Fig. 1c, f, Fig. S3b). PCA of all microRNA samples shows separation of brain samples from other tissues and cell lines by PC1, while cell lines are separated by both PC1 and PC2 (Fig. S2b). Comparison of microRNA detection between tissue types and cell lines reveals that brain samples have the least diversity when compared to the full set of GENCODE v29 microRNAs (Fig. S2c), yet the most microRNAs expressed at ≥ 2 CPM (Fig. S2d). This indicates that a core set of microRNAs are expressed in the brain. Their high tissue-specific expression may be driving the clustering of brain samples apart from non-brain tissues and cell lines, which overlap slightly (Fig. S2b). Comparison of the overlap of detected microRNAs across the sample biotypes reveals that more microRNAs overlap between brain and cell lines than between brain and non-brain tissues (Fig. S2e). Of the 80 shared microRNAs between brain and cell lines, most (53) are expressed in neuronal and glial derived cells.

Machine learning models predict the support for long-read TSS peaks by other TSS-annotating assays and in a cross-cell type manner. We sought to identify a set of high-confidence TSS regions from our observed LR-RNA-seq TSSs using multiple orthogonal TSS assays such as RAMPAGE and CAGE^{40,41}. However, matching data from external assays is only available for a few samples, such as our ENCODE tier 1 cell lines GM12878 and K562. Therefore, we wanted to predict the external support for our observed LR-RNA-seq TSSs. The majority of our observed TSS regions are supported by these external assays (Fig. S6a-b, Fig. 2b, Fig. S5a). We used a simple logistic regression model that incorporates expression, DNase-Hypersensitivity (DHS)⁴², and length of our LR-RNA-seq observed TSSs (Supplementary methods, Fig. S6c-d). Models trained and tested on one experiment each from GM12878 and K562 TSS regions were able to predict whether an LR-RNA-seq TSS region was also supported by RAMPAGE or CAGE assays, with AUROC values as high as 0.95 for Cerberus and 0.98 for LAPA-annotated peaks in the same cell type (Fig. S6e), which is expected given that LAPA regions are narrower than Cerberus regions. Models trained on one cell type can also be used to predict the RAMPAGE or CAGE support in another cell type, in a cross-cell type manner (Fig. S6f). This approach may be used to define a set of high-confidence TSS regions from LR-RNA-seq that would also be supported by RAMPAGE or CAGE where neither RAMPAGE nor CAGE data are available in the cell type of interest. This demonstrates that TSSs derived from LR-RNA-seq serve as a reasonable stand-in for CAGE and RAMPAGE, with the added benefit that LR-RNA-seq profiles both ends and the exon structure of transcripts at the same time.

Applying Cerberus to the human ENCODE4 LR-RNA-seq dataset leads to the largest number of detected alternative splicing events to date. We compared the detection of alternative splicing (AS) events in our dataset with a recent LR-RNA-seq transcriptome published by the GTEx consortium²³. We ran SUPPA2⁴³ on the observed LR-RNA-seq transcripts and obtained, for every gene and type of local AS event, a list of AS transcripts. We found a considerably larger number of AS transcripts compared to those reported in the GTEx LR-RNA-seq catalog. We observed a higher proportion of novel AS transcripts defined by EC compared to TSS and TES (Fig. S8a), albeit lower than those reported by GTEx. This is likely due to the fact that our novel transcripts are defined with respect to a more recent GENCODE version (v40) than the one used by the GTEx study (v26). In support of this, we found that the majority of our observed ENCODE LR-RNA-seq transcripts, both known and novel, are missing in the GTEx catalog (Fig. S8b). On the other hand, although most of the GTEx novel transcripts are not reported in the ENCODE4 catalog, they represent combinations of already annotated splice junctions (NIC). From a methodological perspective, we also found that Cerberus accounts for a larger variety of AS events related to TSSs and TESs ($0.25 < \text{PSI} < 0.75$) compared to SUPPA2 (Fig. S8c). Altogether, this indicates that the ENCODE4 LR-RNA-seq catalog provides the largest set of novel and annotated AS events in the human transcriptome available to date.

SUPPLEMENTARY METHODS

B6/Cast mouse tissue collection. Mouse tissues were harvested from C57BL/6J (RRID:IMSR_JAX:000664) x CAST/EiJ (RRID:IMSR_JAX:000928) F1 animals across 7 postnatal day (PND) or postnatal month (PNM) timepoints: PND4, PND10, PND14, PND25, PND36, 2 months and 18-20 months. Tissues were flash frozen in liquid nitrogen and stored at -80C prior to processing.

Preprocessing short-read RNA-seq data and data availability. All short-read RNA-seq data was preprocessed according to the details on the ENCODE portal. Gene quantification of 548 short RNA-seq datasets were downloaded from the ENCODE portal using this cart (<https://www.encodeproject.org/carts/4ea7a43f-e564-4656-a0de-b09c92215e52/>), then TPM values for polyA genes were extracted from each of them.

Preprocessing microRNA-seq data and data availability. Quantification of 254 microRNA-seq datasets using GENCODE GRCh38 V29 annotations were downloaded from the ENCODE portal using this cart (https://www.encodeproject.org/carts/human_mirna/). Counts were concatenated across all datasets and converted to CPM for downstream analyses.

Preprocessing LR-RNA-seq data and data availability. All LR-RNA-seq data was preprocessed according to the details on the ENCODE portal. Input and output files, including the final Cerberus GTFs, gene triplets, and transcript triplets, are available at the following accessions:

- Human: ENCSR957LMA
- Mouse: ENCSR110KDI

Raw data are available at the following links:

- Human: <https://www.encodeproject.org/carts/829d339c-913c-4773-8001-80130796a367/>
- Mouse: <https://www.encodeproject.org/carts/55367842-f225-45cf-bfbe-5ba5e4182768/>

Human / Mouse LR-RNA-seq annotation with TALON and LAPA. Mapped LR-RNA-seq BAMs were obtained from the ENCODE portal using the above cart links for human and mouse respectively. Reads were annotated with their 3' end A content using the **talon_label_reads** module and hg38 / mm10. Reads were annotated using talon with reference annotation GENCODE v29 / vM21. Output transcripts were filtered for reproducibility of 5 reads across 2 libraries, and for reads that had fewer than 50% A nucleotides in the last 20 bp of the 3' end to remove artifacts of internal priming using the **talon_filter_transcripts** command. Unfiltered and filtered transcript abundance matrices were obtained using the **talon_abundance** command. A filtered GTF was obtained using the **talon_create_GTF** command. From the unfiltered TALON abundance, counts of each gene were computed by summing up counts for each transcript per gene.

We ran LAPA on the bam files mentioned above to create TSS and TES clusters from LR-RNA-seq. If the bam files had replicates, we filtered clusters by choosing a cutoff that ensures a 95% replication rate. Samples without replicates were filtered with a median cutoff of replicated clusters. Using those TSS and TES clusters and the read_annot created by TALON, we corrected TSSs and TESs of the filtered TALON GTF file. During the correction, new transcript isoforms were created if the same exon junction chain mapped to multiple start and end sites.

Gene rank analysis. For detected (≥ 1 TPM in any library) polyA genes in the human LR-RNA-seq dataset, we ranked the genes in each library according to their expression (1 = most highly expressed) and plotted the genes at specific ranks for each library by their TPM, split by cell line and tissue derived libraries. For statistical testing between the cell line and tissue groups, we performed a Wilcoxon rank-sum test with p-value thresholds $P > 0.05$; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, **** $P \leq 0.0001$.

Novel gene analysis. For novel genes in both human and mouse, we first filtered our novel TALON transcripts for those that passed the filters previously described (5 reads in at least 2 libraries and $<50\%$ A nucleotides in the last 20bp of the 3' end). We then selected only the transcripts that passed this filter that belonged to novel intergenic genes and that had at least one spliced (i.e. more than one exon) transcript isoform expressed ≥ 1 TPM. To make an analogous comparison to our annotated genes, we performed the same filtering on our TALON transcripts with the exception of requiring the transcripts to be from annotated polyA genes rather than from novel intergenic genes.

Cerberus overview.

Obtaining annotated TSS / TES regions from GTFs. Given a GTF, **cerberus_gtf_to_bed** (Fig. S4a) will extract the single base pair TSS and TES coordinates and extend them by n bp on either side. Regions within m bp of one another are merged. Each unique combination of coordinates, strand, and gene is recorded in BED format.

Obtaining annotated exon junction chains from GTFs. Given a GTF, **cerberus_gtf_to_ics** will extract each unique combination of intron coordinates, strand, and gene and record them in a tab-separated format (Fig. S4a).

Assigning triplet features numbers. As part of both **cerberus_gtf_to_ends** and **gtf_to_ics** (Fig. S4a), Cerberus numbers triplet features based on their annotation status within the reference GTF, if any. For triplet features derived from these GTFs, each TSS, EC, and TES is numbered from 1 to n within each gene based on the annotation status of the transcript they were derived from. Transcripts are first ordered by MANE status, then APPRIS⁴⁴ principal status, and finally whether the transcript comes from the GENCODE basic set. The result is that triplet features from MANE transcripts are always numbered 1, and lower triplet feature numbers within a gene correspond to transcripts with more importance as determined by GENCODE.

Merging TSS and TES regions across multiple BED files. For each BED file input `cerberus agg_ends` (Fig. S4a) takes a boolean argument for whether the regions should be used to initialize new TSS / TES regions, a boolean argument for whether the regions should be considered reference regions, and a name for each BED file source. BED files without a gene identifier cannot be used to initialize regions. For the first BED file, Cerberus creates a set of reference regions and uses the triplet feature numbers that were previously assigned by Cerberus to name each TSS or TES. The first BED file must have gene IDs and must be used to initialize the regions. For each subsequent BED file, in order, Cerberus determines which new regions are within m bp of a region already in the reference. These regions are added as sources of support for the already-existing regions, but do not extend the boundaries of existing regions in order to combat growing regions as more data is added. If a region is not within m bp of an existing region and the initialize regions option is turned on, the new region is added as a new region in the reference set. After all new regions have been added, triplet feature numbers are computed by ordering the features within each gene based on the number assigned by Cerberus in a previous step and then incrementing the preexisting Cerberus reference maximum number. BED files that are not used to initialize new regions will only ever be added as additional forms of support for each region already in the reference.

Merging ECs across multiple EC files. For each EC file, `cerberus agg_ics` (Fig. S4a) takes a boolean argument for whether the ECs should be used as a reference and a source name. For the first EC file, Cerberus creates a reference set of ECs and uses the EC numbers that were determined using `cerberus gtf_to_ics`. For each subsequent EC file, Cerberus finds ECs that are not already in the Cerberus reference set, orders the new ECs by their numbers from `cerberus gtf_to_ics`, creates new numbers for each EC by, in order, assigning them numbers by incrementing from the maximum existing number for a gene from the reference.

Creating a Cerberus reference. After generating an aggregated TSS, EC, and TES file, `cerberus write_reference` (Fig. S4a) will write all three tables in a Cerberus reference h5 format, a well-supported and commonly used data structure that can store multiple tables.

Updating a GTFs and counts matrices with a Cerberus annotation. After each transcript from a transcriptome has been assigned a transcript triplet, the corresponding GTF and counts matrix from the transcriptome can be updated to use the new transcript identifier using `cerberus replace_gtf_ids` and `cerberus replace_ab_ids` (Fig. S4b). Cerberus will replace the transcript ids with the transcript triplets and, if requested, merge transcripts that are assigned duplicate transcript triplets, summing the counts in the case of the counts matrix.

Gene triplet and gene structure simplex coordinate computations. Following transcriptome annotation, gene triplets can be calculated for different sets of annotated transcripts using Cerberus' Python API and the `CerberusAnnotation data structure`. Regardless of the input set, Cerberus computes the gene triplets by counting the number of unique TSSs, ECs, and TESs used across a set of transcripts (Fig. 3a, Fig. S9). This calculation can be done without any filtering using the `CerberusAnnotation.get_source_triplets()` function, which computes the number of TSSs, ECs, and TESs used across each transcriptome annotated by Cerberus. `CerberusAnnotation.get_expressed_triplets()` will calculate the gene triplets for individual samples based on the subset of transcripts that are expressed in each sample and can optionally use a table of transcript / sample combinations to determine which transcripts are used in each sample. Finally, `CerberusAnnotation.get_subset_triplets()` simply takes in a list of transcripts to compute a gene triplet for the entire input set. In all cases, the number of transcripts used to calculate the gene triplet is also recorded. After computing the gene triplets, the EC count is converted to the splicing ratio. To generate the gene structure simplex coordinates, the sum of the number of TSSs, splicing ratio, and number of TESs is normalized such that they sum to one (Fig. 3a-b, Fig. S9).

Additionally, the sector assignments are generated for each gene triplet. Genes with a TSS simplex coordinate >0.5 are TSS-high, those with a TES simplex coordinate >0.5 are TES-high, and those with a splicing ratio simplex coordinate >0.5 are splicing-high. Genes where all three simplex coordinates ≤ 0.5 are mixed, and genes with just one transcript are in the simple sector. An important note is that mixed genes can have the same coordinates as a simple gene. To this end, when calculating gene triplets, the number of transcripts used to generate the triplet is also recorded and used to separate out the simple from the mixed genes (Fig. 3a-b, Fig. S9).

Computing gene triplet centroids. Given a set of gene triplets, the centroid is computed by averaging each gene structure simplex coordinate. The resulting coordinate retains the property that it sums to one (Fig. S9).

Computing distances in the gene structure simplex. We compute the distance between any two points on the gene structure simplex as the Jensen-Shannon distance (Fig. S9). Jensen-Shannon distance is a metric on probability distributions⁴⁵. For a given pair of gene structure simplex coordinates, the Jensen-Shannon distance is computed in Cerberus using Scipy⁴⁶ with the `scipy.spatial.distance.jensenshannon` function.

Cerberus processing of human ENCODE4 LR-RNA-seq dataset.

Obtaining annotated TSS / TES regions from GTFs. The GTF files from GENCODE v40, GENCODE v29, the LAPA output GTF representing the human ENCODE LR-RNA-seq dataset, and the GTEx LR-RNA-seq GTF were used to obtain TSS and TES regions associated with each transcript using `cerberus gtf_to_ends`. For each GTF, the single base pair TSS and TES coordinates were extracted and extended 50 bp on either side, and regions within 50 bp of one another were merged. Each unique combination of coordinates, strand, and gene were recorded.

Obtaining external TSS / TES regions. External datasets used to support TSSs were obtained from the ENCODE CAGE and RAMPAGE data, FANTOM CAGE data4, and ENCODE PLS, pELS, and dELS cCREs. External datasets used to support TESs were obtained from ENCODE PAS-seq data, and the PolyA Atlas¹⁸. Each file was downloaded in BED format and converted to the BED format required for Cerberus.

Obtaining annotated exon junction chains from GTFs . The GTF files from GENCODE v40, GENCODE v29, from the human ENCODE LR-RNA-seq output GTF, and the GTEx LR-RNA-seq GTF were used to obtain exon junction chains from each transcript using `cerberus gtf_to_ics`. Each unique combination of intron coordinates, strand, and gene were recorded.

Creating a set of reference triplet features. To create a consensus reference set of triplet features, `cerberus agg_ends` and `cerberus agg_ics` (Fig. S4a) were run on the aforementioned TSS, EC, and TES sets, with $m=20$ for the TSSs and TESs. The triplet features from GENCODE v40 and v29 were used as reference features. For the TSSs, new regions were incorporated from GENCODE v40, v29, the human ENCODE LR-RNA-seq data, and the GTEx data, whereas the CAGE, RAMPAGE, and cCRE data were only used as forms of support for existing regions. For TESs, new regions were incorporated from GENCODE v40, v29, the human ENCODE LR-RNA-seq data, and the GTEx data, whereas the PAS-seq and PolyA Atlas regions were used as forms of support for existing regions.

Transcriptome annotation. The GTFs of the GENCODE v40, GENCODE v29, and human ENCODE LR-RNA-seq transcriptomes were annotated with `cerberus annotate_transcriptome`, updated GTFs were generated with `cerberus replace_gtf_ids` with the update ends and collapse options used (Fig. S4b). For the human ENCODE LR-RNA-seq data, `cerberus replace_ab_ids` was also run on the filtered abundance file output from LAPA using the collapse option to generate a matching counts matrix (Fig. S4b).

Cerberus analysis of human ENCODE4 LR-RNA-seq.

Finding observed transcripts and transcripts expressed in a sample. Observed transcripts are defined as transcripts that are expressed ≥ 1 TPM in any given library. Observed transcripts in a specific sample are transcripts that are expressed ≥ 1 TPM in any library that belongs to the same sample.

Finding observed major transcripts and major transcripts in a sample. For each sample, each transcript is assigned a percent isoform (π , 0-100) value that indicates what percentage of the gene's expression is derived from said transcript using Swan⁴⁷. Transcripts for a gene are then ranked by π value. In order from the highest π value transcript to the lowest π value transcript, transcripts are added to the major transcript set until the cumulative π value of the set is >90 , yielding the sample-level major transcript set. The observed major transcripts for the entire dataset is computed by taking the union of all major transcripts across all samples. In both cases, transcripts are limited to those that have passed the observed and sample-level observed transcripts as defined above.

Gene triplet computations. Gene triplets were calculated for the following sets of transcripts, all just using polyA genes:

- All transcripts from annotated GENCODE v40 genes (v40)
- All observed transcripts (observed)
- All observed major transcripts (observed major)
- Detected transcripts in each sample (sample-level)
- Detected major transcripts in each sample (sample-level major)
- All observed transcripts in the dataset from samples that match the mouse samples (mouse match)
- All observed major transcripts in the dataset from samples that match the mouse samples (mouse match major)

Transcriptional diversity by gene biotype comparison. Using the gene triplets table, we found the gene / sample combination where each polyA gene is most highly expressed and recorded the gene TPM and number of transcripts from that gene in that sample. We then split each gene into its biotype category (protein coding, lncRNA, or pseudogene) and into a TPM bin (lowly expressed, 1-10 TPM; medium expressed, 10-100 TPM; and highly expressed, 100-max TPM).

Gene structure simplex distances computed. We computed the follow pairwise distances between simplex points:

- Sample-level gene triplet vs. the centroid for the sample-level gene triplets
- Observed gene triplet vs. the centroid for the sample-level gene triplets for each gene with at least 2 transcripts
- Observed gene triplet vs. observed major gene triplet

Each set of distances was computed using only protein coding genes. Z-scores were also computed for each comparison.

Comparing sample-level to observed gene triplets. The number of transcripts, TSSs, ECs, and TESs was calculated for each gene globally (i.e. transcripts or triplet features / gene) and for each sample (i.e. transcripts or triplet features / gene / sample). For transcripts, TSSs, ECs, and TESs separately, a two-sided KS test was performed using Scipy's `stats.kstest` function to assess statistical differences between the global and sample-level transcripts or triplet features per gene distributions.

Calling predominant transcripts. On both the sample and library level, we called the most highly expressed transcript from a gene the predominant transcript for that gene. On the sample level, we used the mean expression of the transcript.

Predominant transcript MANE comparison. We first restricted this analysis to only consider genes which have annotated MANE transcripts. For these genes, we determined how often the predominant transcript for a given gene is the MANE transcript for a gene in each library.

Cerberus processing of mouse ENCODE4 LR-RNA-seq dataset.

Obtaining annotated TSS / TES regions from GTFs. The GTF files from GENCODE vM25, GENCODE vM21, and from the LAPA output GTF representing the mouse ENCODE LR-RNA-seq dataset were used to obtain TSS and TES regions associated with each transcript using `cerberus gtf_to_ends`. For each GTF, the single base pair TSS and TES coordinates were extracted and extended 50 bp on either side, and regions within 50 bp of one another were merged. Each unique combination of coordinates, strand, and gene were recorded.

Obtaining external TSS / TES regions. External datasets used to support TSSs were obtained from the ENCODE mouse PLS, pELS, and dELS cCREs. External datasets used to support TESs were obtained from the mouse PolyA Atlas. Each file was downloaded in BED format and converted to the BED format required for Cerberus.

Obtaining annotated exon junction chains from GTFs. The GTF files from GENCODE vM25, GENCODE vM21, and from the mouse LR-RNA-seq GTF were used to obtain exon junction chains from each transcript using `cerberus gtf_to_ics`. Each unique combination of intron coordinates, strand, and gene were recorded.

Creating a set of reference triplet features. To create a consensus reference set of triplet features, `cerberus agg_ends` and `cerberus agg_ics` (Fig. S4a) were run on the aforementioned TSS, EC, and TES sets, with $m=20$ for the TSSs and TESs. The triplet features from GENCODE vM25 and vM21 were used as reference features. New TSSs were incorporated from GENCODE vM25, vM21, the mouse ENCODE LR-RNA-seq data, whereas the cCRE data were only used as forms of support for existing regions. New TESs were incorporated from GENCODE vM25, vM21, and the mouse ENCODE LR-RNA-seq data, whereas the PolyA Atlas regions were used as forms of support for existing regions.

Transcriptome annotation. The GTFs of the GENCODE vM25, GENCODE vM21, and mouse ENCODE LR-RNA-seq transcriptomes were annotated with `cerberus annotate_transcriptome`, updated GTFs were generated with `cerberus replace_gtf_ids` with the update ends and collapse options used (Fig. S4b). For the mouse ENCODE LR-RNA-seq data, `cerberus replace_ab_ids` was also run on the filtered abundance file output from LAPA using the collapse option to generate a matching counts matrix (Fig. S4b).

Cerberus analysis of mouse ENCODE4 LR-RNA-seq.

Finding observed transcripts and transcripts expressed in a sample. Observed transcripts are defined as transcripts that are expressed ≥ 1 TPM in any given library. Observed transcripts in a specific sample are transcripts that are expressed ≥ 1 TPM in any library that belongs to the same sample.

Finding observed major transcripts and major transcripts in a sample. For each sample, each transcript is assigned a percent isoform (pi, 0-100) value that indicates what percentage of the gene's expression is derived from said transcript using Swan⁴⁷. Transcripts for a gene are then ranked by pi value. In order from the highest pi value transcript to the lowest pi value transcript, transcripts are added to the major transcript set until the cumulative pi value of the set is >90, yielding the sample-level major transcript set. The observed major transcripts for the entire dataset is computed by taking the union of all major transcripts across all samples. In both cases, transcripts are limited to those that have passed the observed and sample-level observed transcripts as defined above.

Gene triplet computations. Gene triplets were calculated for the following sets of transcripts; all just using polyA genes:

- All annotated GENCODE vM25 genes (vM25)
- All observed transcripts in the dataset (observed)
- All observed major transcripts in the dataset (observed major)
- Detected transcripts in each sample (sample-level)
- Detected major transcripts in each sample (sample-level major)

Calling predominant transcripts. On both the sample and library level, we called the most highly expressed transcript from a gene the predominant transcript for that gene. On the sample level, we used the mean expression of the transcript.

Human-mouse comparison. We found orthologous genes between human and mouse using [this Biomart query](#), and subset our considered genes to those that were protein coding, expressed in both species, and were just 1:1 orthologs. We determined the sector of each gene in each species using the observed major gene triplets in mouse, and the mouse match major gene triplets in human. We counted the number of genes that have the same sector between human and mouse. Furthermore, we compared the sector of each orthologous pair of genes between species just in the matching embryonic stem cell samples (H1 in human, F121-9 in mouse) between human and mouse to verify that the trend seen overall was reproducible on a more one to one comparison. Additionally, we computed the centroids from the sample-level gene triplets from matching samples in human and all sample-level gene triplets in mouse mouse and calculated the Jensen-Shannon distances between sample-level centroids for each orthologous gene in human and mouse.

ORF and NMD prediction. We used TAMA's⁴⁸ ORF / NMD prediction pipeline with [minimal changes](#) to support our file formats. To pick one representative ORF from each transcript, we chose the ORF with the highest percent identity from BLASTP⁴⁹ to an annotated GENCODE v40 protein sequence; breaking ties by considering ORF completeness. For transcripts with no BLASTP hits to known transcripts, we picked complete ORFs; breaking ties by picking the longest ORF.

Comparing detection of AS events by SUPPA and Cerberus. We used SUPPA2 (v2.3⁴³) to define alternative splicing (AS) events (A3: alternative 3' splicing; A5: alternative 5' splicing; AF: first exon; AL: last exon; IR: intron retention; SE: exon skipping; MX: mutually exclusive exons). Specifically, we generated a catalog of local AS events based on the Cerberus GTF file (function `generateEvents`) and used the novelties of each observed transcript to compute the proportion of novel transcripts (based on EC, TSS, or TES) out of the total set of transcripts involved in a particular type of event.

Next, we used SUPPA2 to compute the Proportion of Splicing Index (PSI) for each type of event using the observed transcript filtered expression matrix (polyA transcripts expressed ≥ 1 TPM in at least one library; function `psiPerEvent`). PSI values were averaged between replicates of the same sample. We selected genes with at least one local AF or AL event, applying a threshold of $0.25 < \text{PSI} < 0.75$.

In order to compare the detection of AS events by SUPPA2 and Cerberus, we also computed triplet feature PSI values based on events identified by Cerberus by dividing the counts for any given TSS, EC, or TES by the total counts for the gene in a given sample. We selected genes with at least one local event at the TSS or TES ($0.25 < \text{PSI} < 0.75$). Next, we computed the intersection between genes showing AF (SUPPA2) and TSS (Cerberus) events, and between genes showing AL (SUPPA2) and TES (Cerberus) events.

Machine learning models for RAMPAGE and CAGE TSS prediction. RAMPAGE and CAGE TSS annotation data for GM12878 and K562 were obtained from ENCODE portal (ENCSR000AEI, ENCSR000AER, ENCSR000CJN, ENCSR000CKA). LAPA and Cerberus TSS regions derived from just one experiment each for GM12878 and K562 (ENCSR962BVU and ENCSR589FUJ respectively) were used for long-read data. Using bedtools intersect⁵⁰ a binary (0/1) label for each long-read peak was assigned depending on whether the region overlapped with at least one peak in either of the RAMPAGE or CAGE assays in the same cell type. Average DHS signal values over LR TSS peaks were calculated using UCSC `bigWigAverageOverBed` on GM12878 and K562 DHS-seq experiments (ENCSR000EMT, ENCSR000EOT). Test sets include long-read regions from chromosomes 2 and 3, whereas training sets include all other human chromosomes. 7 logistic regression models were trained on each long-read experiment using all the $2^3 - 1$ combinations of peak's TPM expression,

DHS signal, and length (i.e. in R: `glm(label ~ TPM + DHS + length, type = "binomial")`) where the input parameters have been log₂-transformed) and the AIC values were calculated and ranked in each experiment and for each model type. The model using all 3 parameters (TPM, DHS, and length) had the lowest AIC, meaning that given the number of parameters and the observed RSS error, `logit[label ~ TPM + DHS + length]` had the highest predictive power and was therefore selected. For the same cell-type prediction, a model is trained on [chr1, chr4-22, chrX] and tested on [chr2, chr3] for long-read data from the same cell type (ex: K562). In cross-cell type prediction, a model is trained and tested on two different cell lines (ex: trained on [chr1, chr4-22, chrX] of a GM12878 long-read experiment and tested on [chr2, chr3] of a K562 long-read experiment). Cerberus replicates belonging to the same experiment were combined by taking the average mean-normalized TPM values of the identical peaks across different replicates.

Data and code availability.

- Human LR-RNA-seq data / processing pipeline: <https://www.encodeproject.org/annotations/ENCSR957LMA/>
- Mouse LR-RNA-seq data / processing pipeline: <https://www.encodeproject.org/annotations/ENCSR110KDI/>
- Processing / figure generation code: https://github.com/fairliereese/paper_rnawg
- Cerberus: <https://github.com/fairliereese/cerberus>

SUPPLEMENTARY TABLES

- **Table S1: Human LR-RNA-seq library metadata.**
- **Table S2: Mouse LR-RNA-seq library metadata.**

SUPPLEMENTARY FIGURES

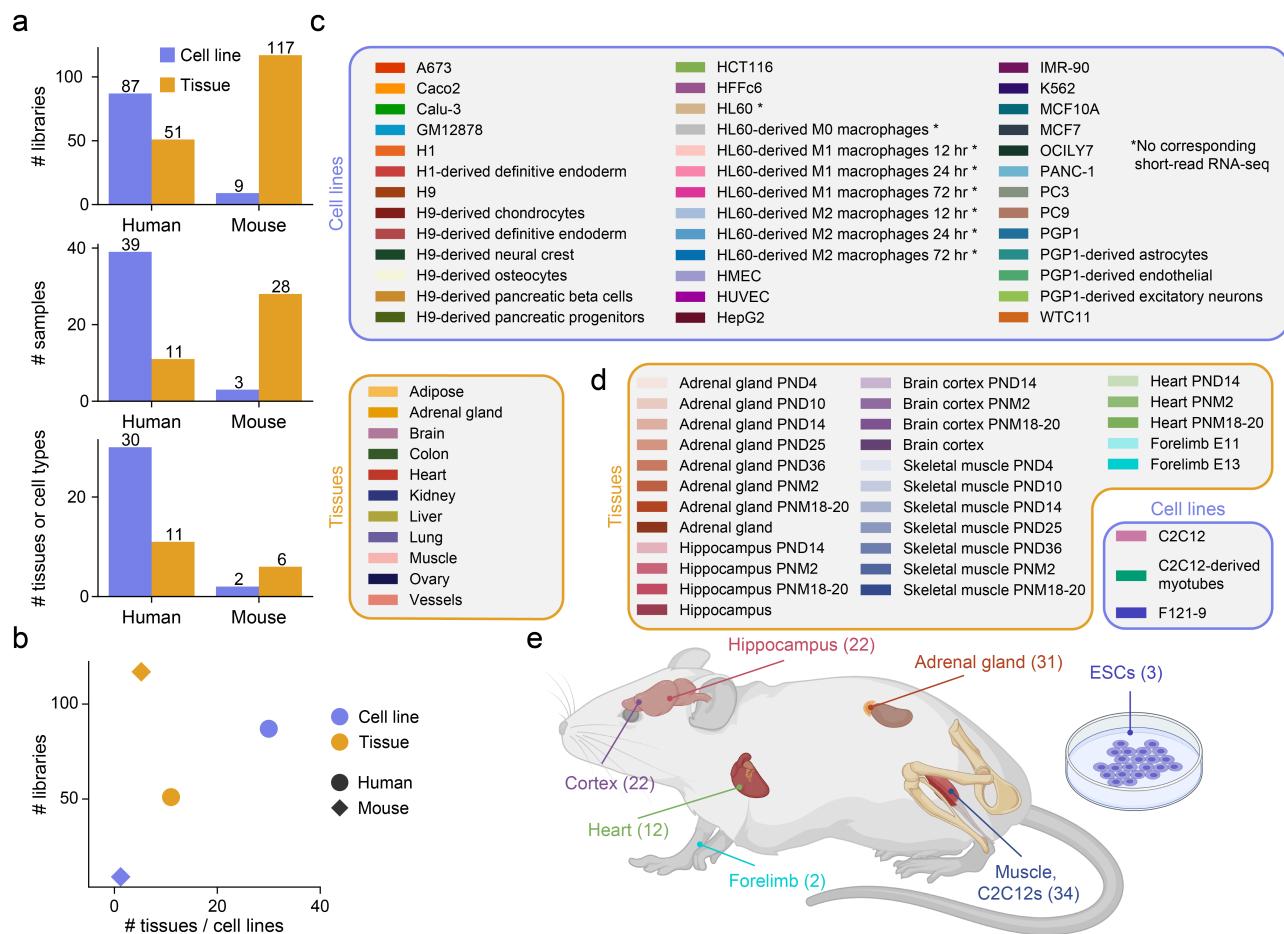


Figure S1. Overview of the ENCODE4 LR-RNA-seq dataset. **a**, From top to bottom, number of LR-RNA-seq libraries, samples (split by cell line / tissue identity as well as timepoint, when relevant), and unique tissues or cell types in the ENCODE LR-RNA-seq dataset split by species and tissue or cell line. **b**, Number of LR-RNA-seq libraries versus the number of tissues or cell lines assayed, split by species and cell line / tissue. **c-d**, Color legend and labels for each **c**, human sample, with samples that lack corresponding short-read RNA-seq data denoted by a star **d**, mouse sample in the LR-RNA-seq dataset; split by tissues and cell lines. **e**, Overview of the sampled tissues and number of libraries from each tissue in the ENCODE mouse LR-RNA-seq dataset.

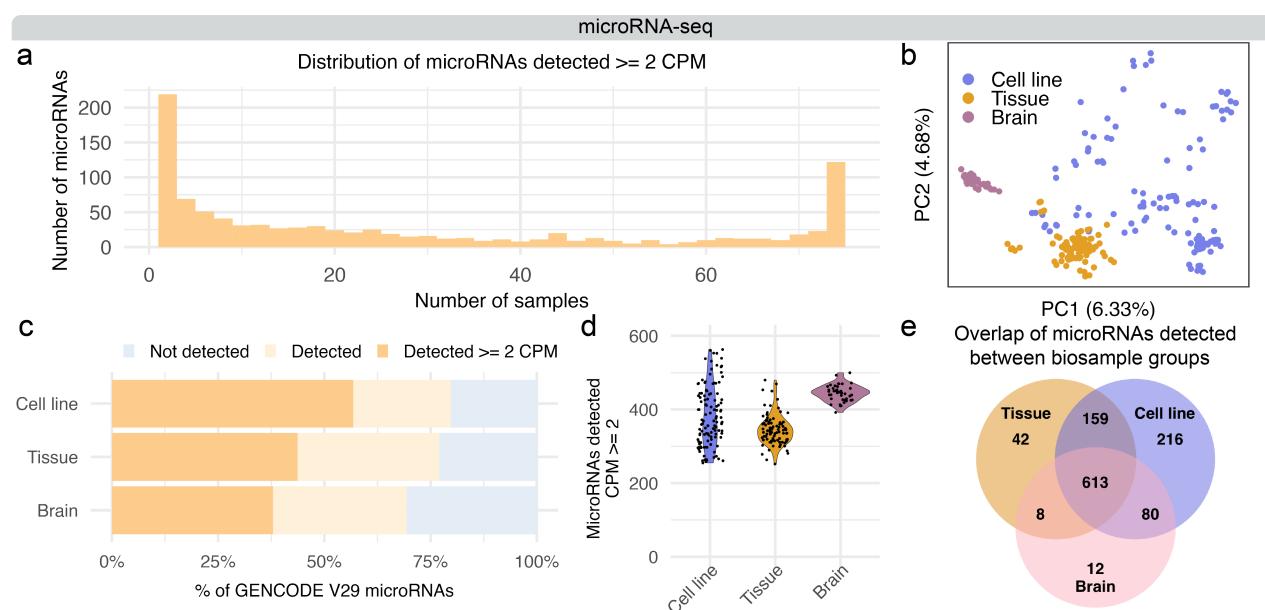


Figure S2. Overview and detection of microRNAs in the ENCODE microRNA-seq dataset. **a**, Distribution of GENCODE v29 mature microRNAs detected at CPM > 2 between cell lines, tissues, and brain tissue samples. **b**, PCA computed on microRNAs detected > 2 CPM in each human microRNA-seq library, colored by cell line and tissue designation and by brain tissue. **c**, Percentage of GENCODE v29 microRNAs detected in at least one ENCODE human microRNA-seq library from either cell line, tissue, or brain tissue samples at > 0 CPM and > 2 CPM. **d**, Number of samples in which each GENCODE v29 microRNA is detected at > 2 CPM in the ENCODE human microRNA-seq dataset. **e**, Overlap of detected (> 2 CPM) microRNAs in at least one library derived from cell line, tissue, or brain tissue.

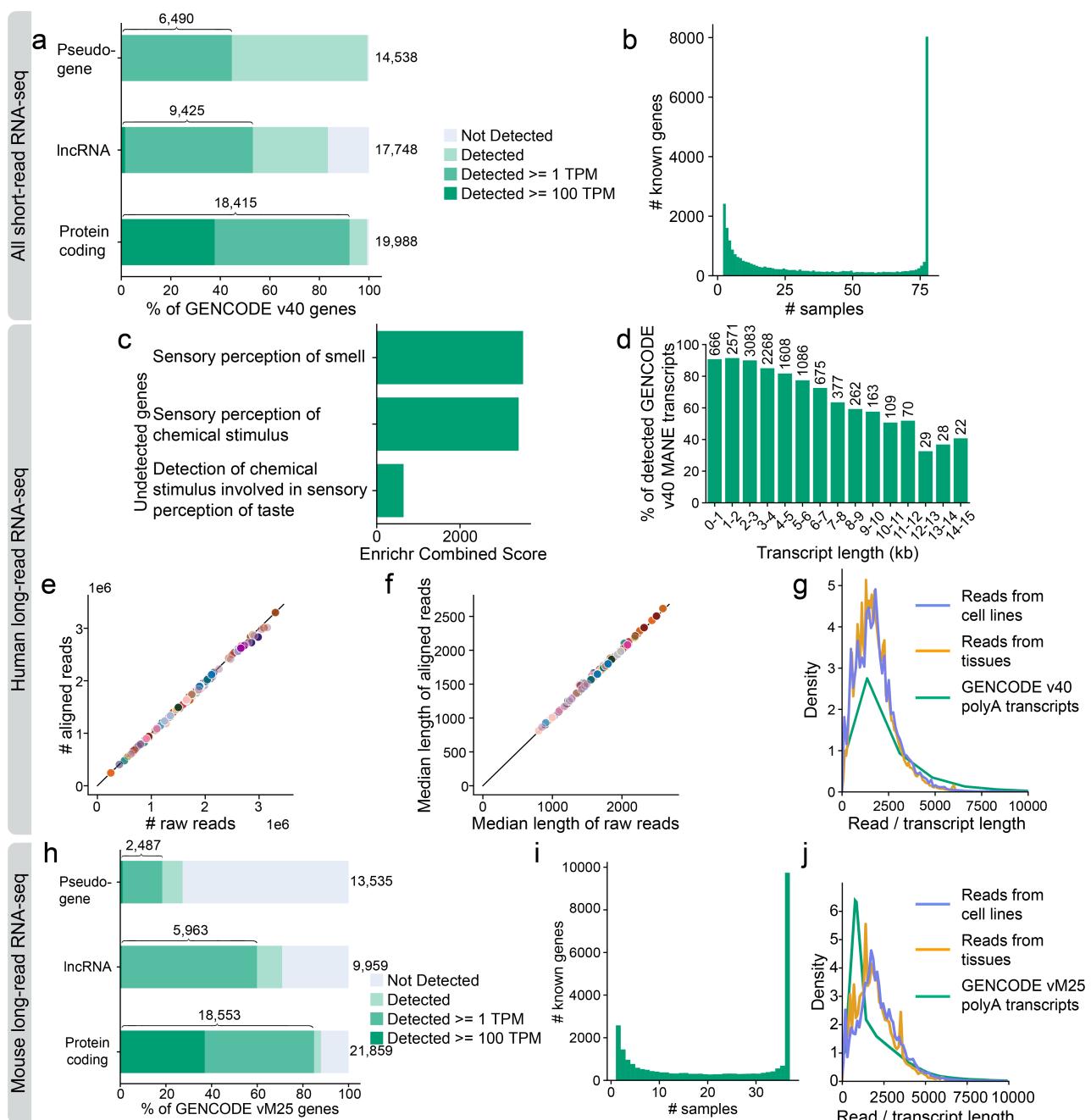


Figure S3. Gene detection from short-read RNA-seq; gene detection, read length and alignment QC in both human and mouse LR-RNA-seq. **a**, Percentage of GENCODE v40 polyA genes by gene biotype detected in at least one ENCODE short-read RNA-seq library from all samples at >0 TPM, >=1 TPM, and >=100 TPM. **b**, Number of samples in which each GENCODE v40 gene is detected >= 1 TPM in the ENCODE short-read RNA-seq dataset in all samples. **c**, Top 3 biological process GO terms from GENCODE v40 protein coding genes that were not detected in the human LR-RNA-seq dataset. **d**, Percentage (y-axis) and number (top of each bar) of GENCODE v40 MANE transcripts we detect binned by transcript length. Restricted to expressed genes that we detect >= 10 TPM in at least one library. **e**, Number of raw reads vs. number of aligned reads in each human LR-RNA-seq library. **f**, Median length of each raw read vs. median length of the aligned portion of each read in each human LR-RNA-seq library. **g**, Read length profiles of post-TALON reads from LR-RNA-seq data split by tissue or cell line designation and polyA transcript length profile from GENCODE v40. **h**, Percentage of GENCODE vM25 polyA genes by gene biotype detected in at least one ENCODE mouse LR-RNA-seq library at >0 TPM, >=1 TPM, and >=100 TPM. **i**, Number of samples that each GENCODE vM25 gene is detected >= 1 TPM in the ENCODE mouse LR-RNA-seq dataset. **j**, Read length profiles of post-TALON reads from mouse LR-RNA-seq data split by tissue or cell line designation and polyA transcript length profile from GENCODE vM25.

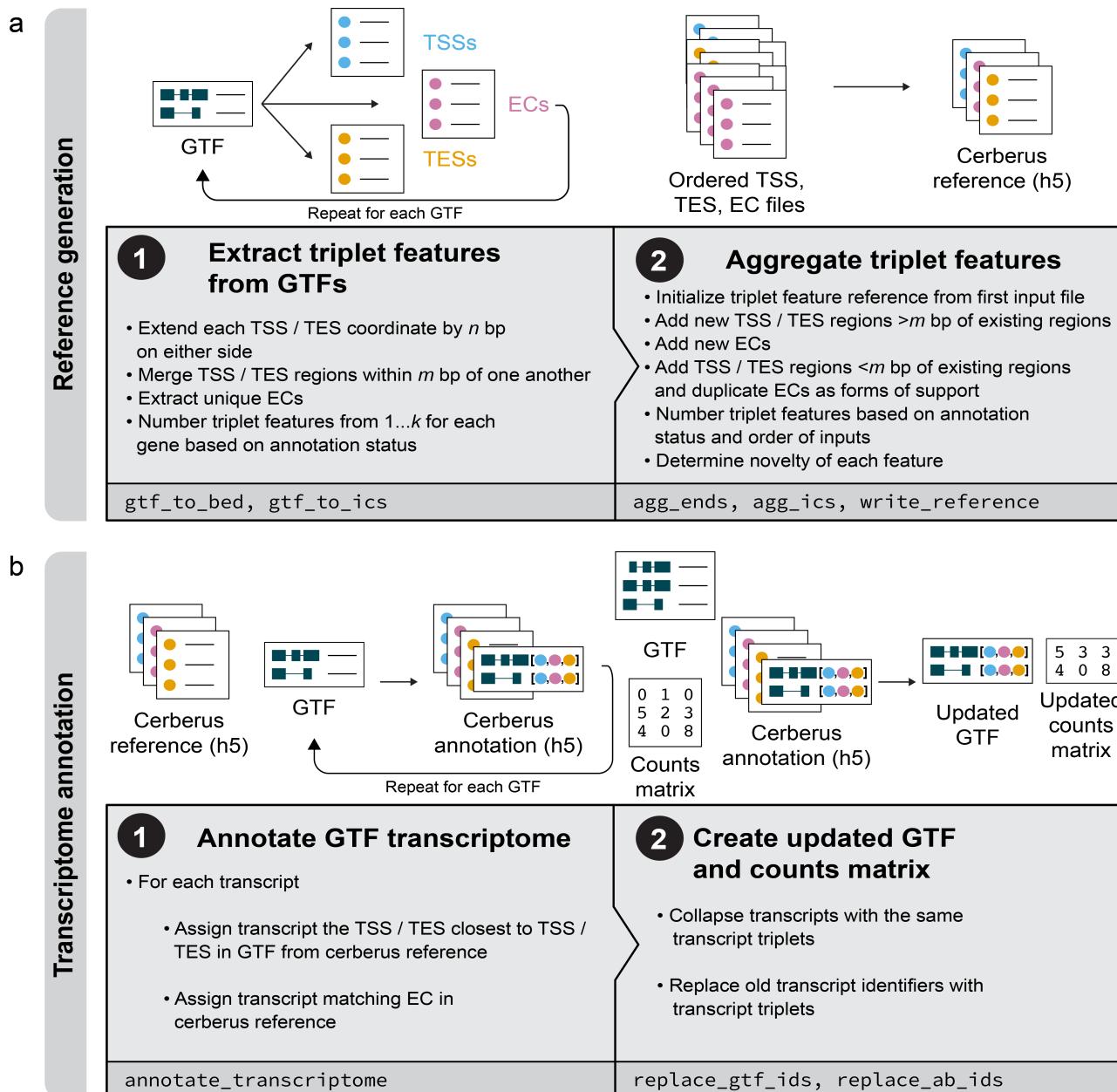


Figure S4. Overview of Cerberus processing of transcriptomes and triplet features. a, Workflow for generating a Cerberus reference: a collection of TSSs, ECs, and TESs (triplet features) sourced from various inputs. **b,** Workflow for generating a Cerberus transcriptome annotation, which assigns each transcript in a GTF a set of triplet features (TSS, EC, TES) from the Cerberus reference.

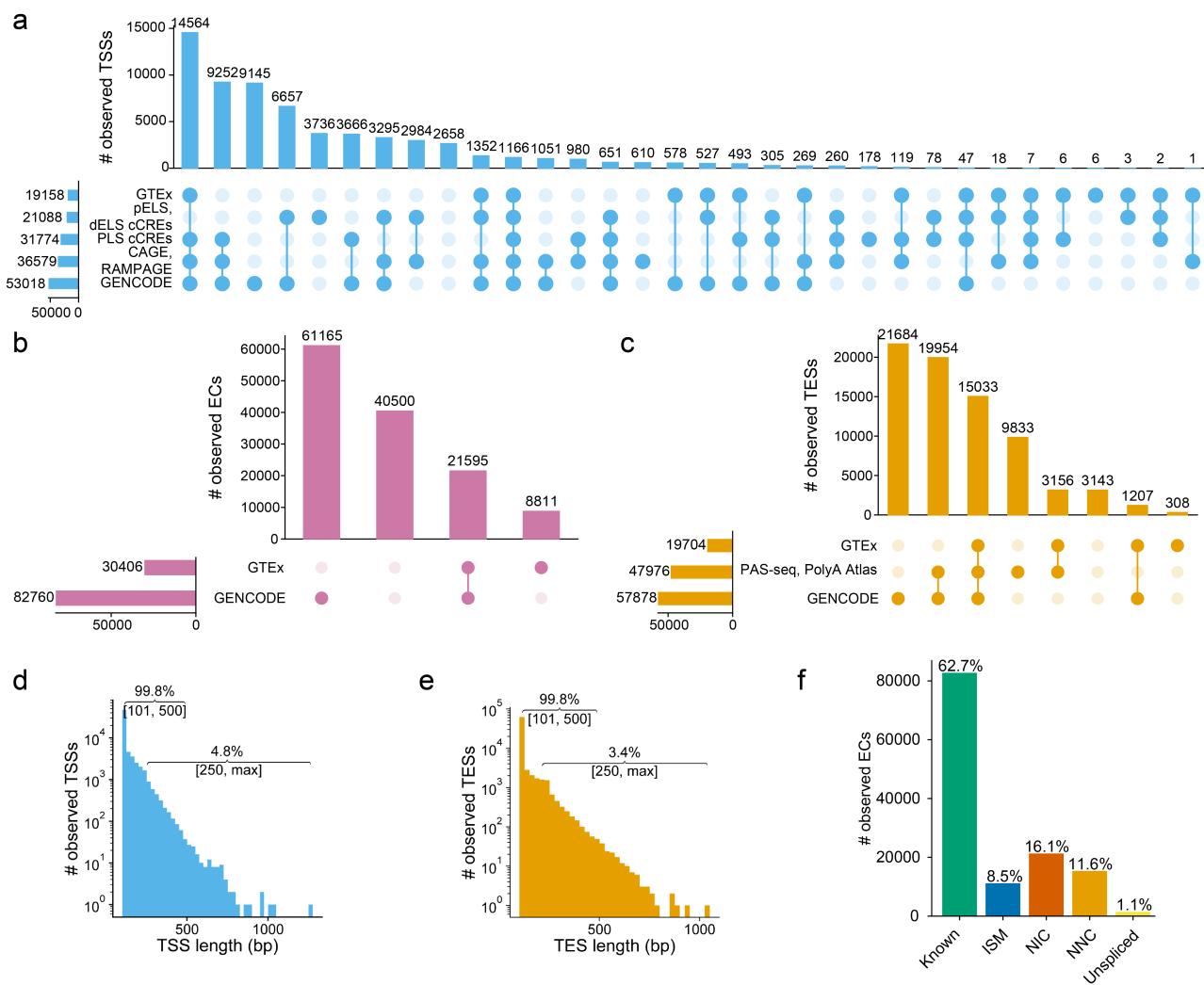


Figure S5. Characterization of observed triplet features from human LR-RNA-seq. **a-c**, Upset plots showing sources that overlap observed triplet features derived from human ENCODE LR-RNA-seq for **a**, TSSs **b**, ECs **c**, TESs. **d-e**, Lengths of observed **d**, TSSs **e**, TESs derived from human ENCODE LR-RNA-seq. **f** Novelty of unique ECs detected ≥ 1 TPM from polyA genes in human ENCODE LR-RNA-seq.

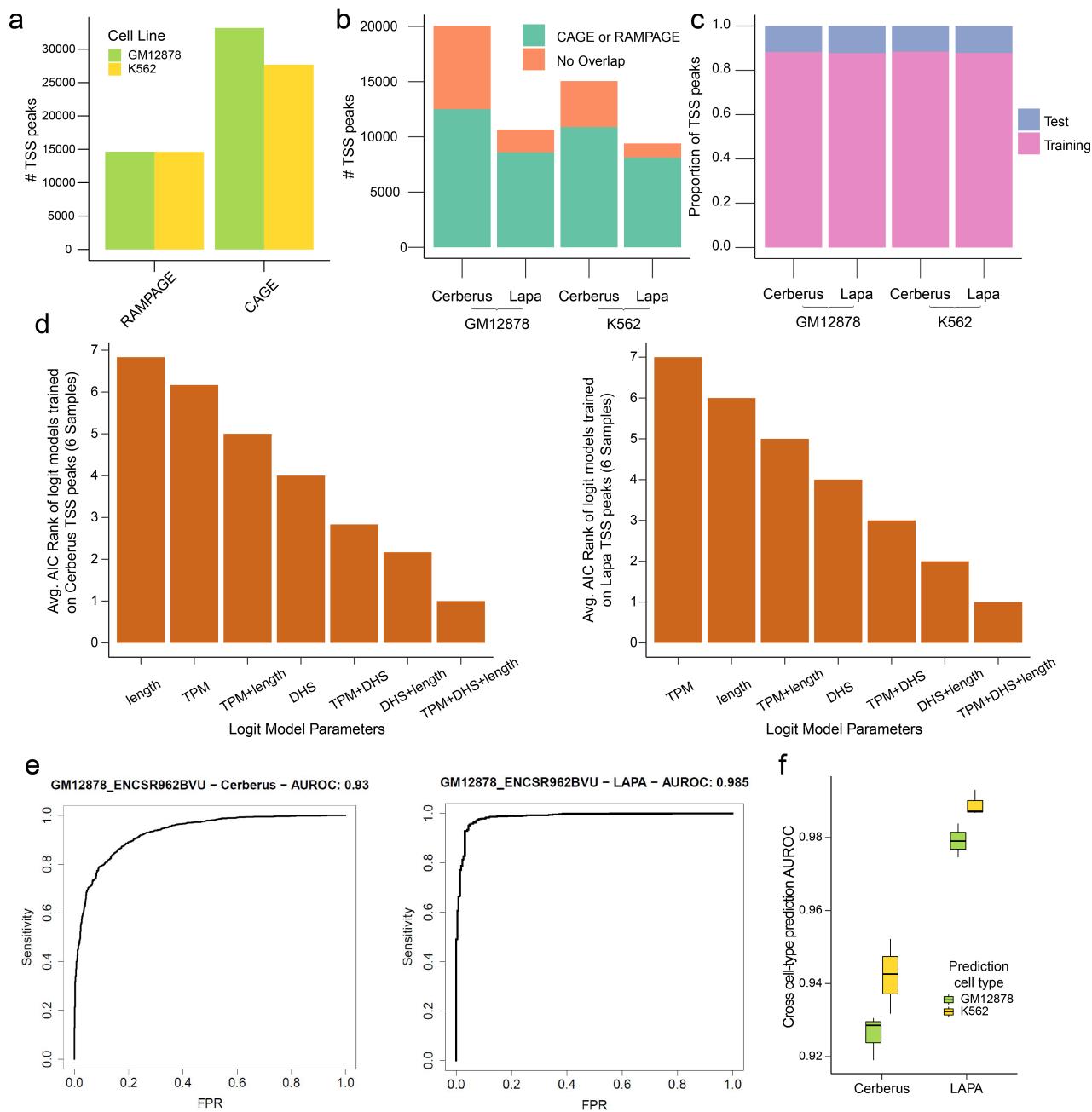


Figure S6. Machine learning models predict support for long-read TSS peaks by other TSS assays in a cross-cell type manner. **a.** Number of long-read RNA-seq TSS peaks called by RAMPAGE and CAGE. **b.** Number of long-read TSS peaks called by Cerberus or LAPA supported by RAMPAGE or CAGE assays in GM12878 and K562 long-read experiments gm12878_3 and k562_2. **c.** Fraction of peaks used for the test (chr 2 and chr3) and training sets (all other chromosomes) in K562 and GM12878 long-read experiments. **d.** Akaike Information Criterion (AIC) values for logistic regression models trained using different sets of parameters. For each experiment, the AIC values for the 7 training settings have been ranked. The y-axis is the average ranking of each model over all GM12878 and K562 long-read TSS peaks from Cerberus (left) and LAPA (right), where logit[overlap TPM + DHS + peak_length] is the best model (i.e. with the lowest AIC). **e.** Same-cell type ROC curves for logit[overlap TPM + DHS + peak_length]. Models tested on chr2 chr3 and trained on other chromosomes in the same cell line. **f.** Cross-cell type logit[overlap TPM + DHS + peak_length]. Distribution of AUROC values for long-read TSS experiments. Ex: To predict if a long-read TSS peak in K562 overlaps with a region in GM12878.

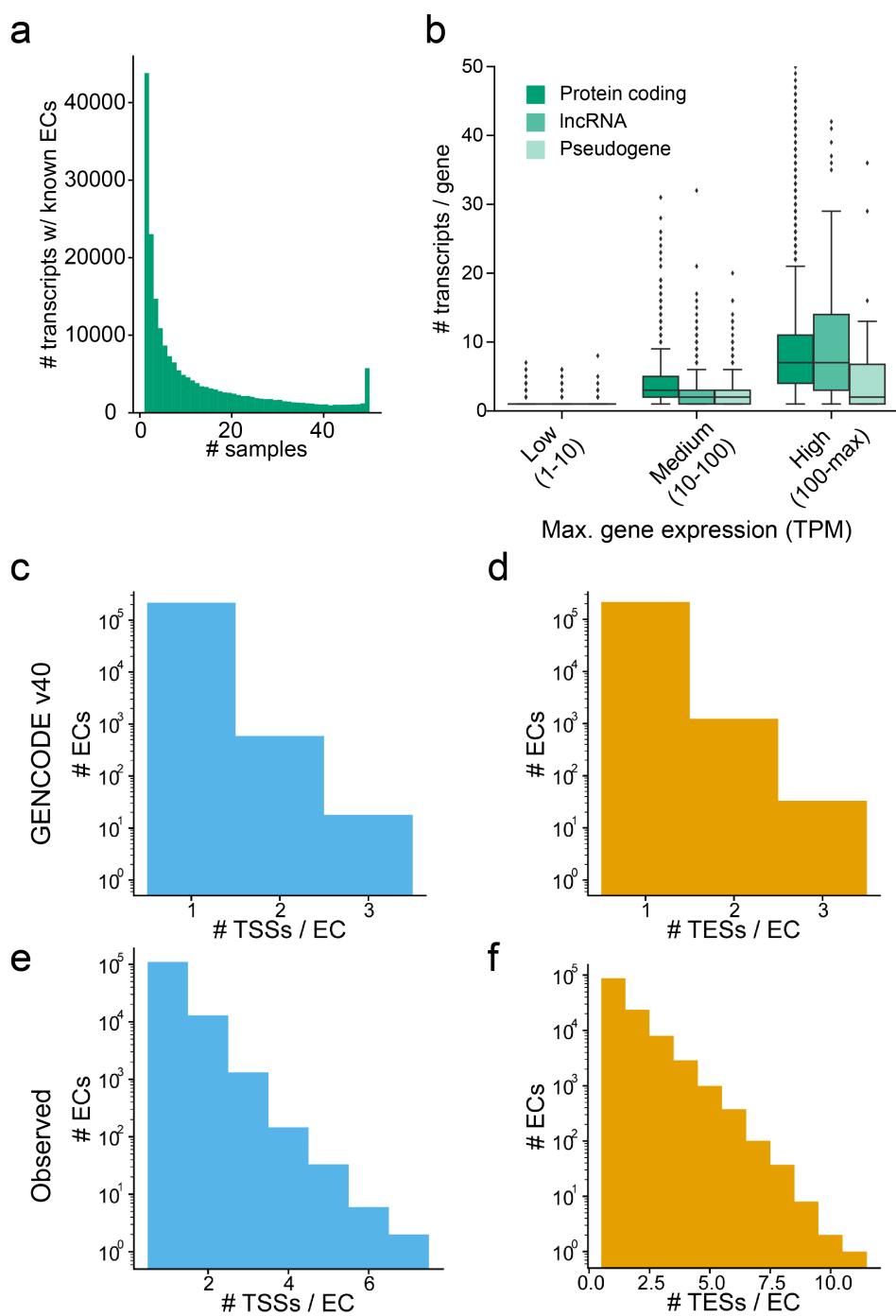


Figure S7. Characterization of observed transcripts from human LR-RNA-seq. **a**, Number of samples in which each transcript with a known EC is detected ≥ 1 TPM in the human ENCODE LR-RNA-seq dataset. **b**, Boxplot of, for the sample where a gene is most highly expressed, the number of transcripts expressed in that sample versus the TPM of the gene in that sample; split by gene biotype and gene expression bin. **c-f**, Number of unique TSSs or TESs per EC with at least 2 exons from transcripts **c-d**, annotated to polyA genes in GENCODE v40, **e-f**, detected ≥ 1 TPM from polyA genes in human ENCODE LR-RNA-seq.

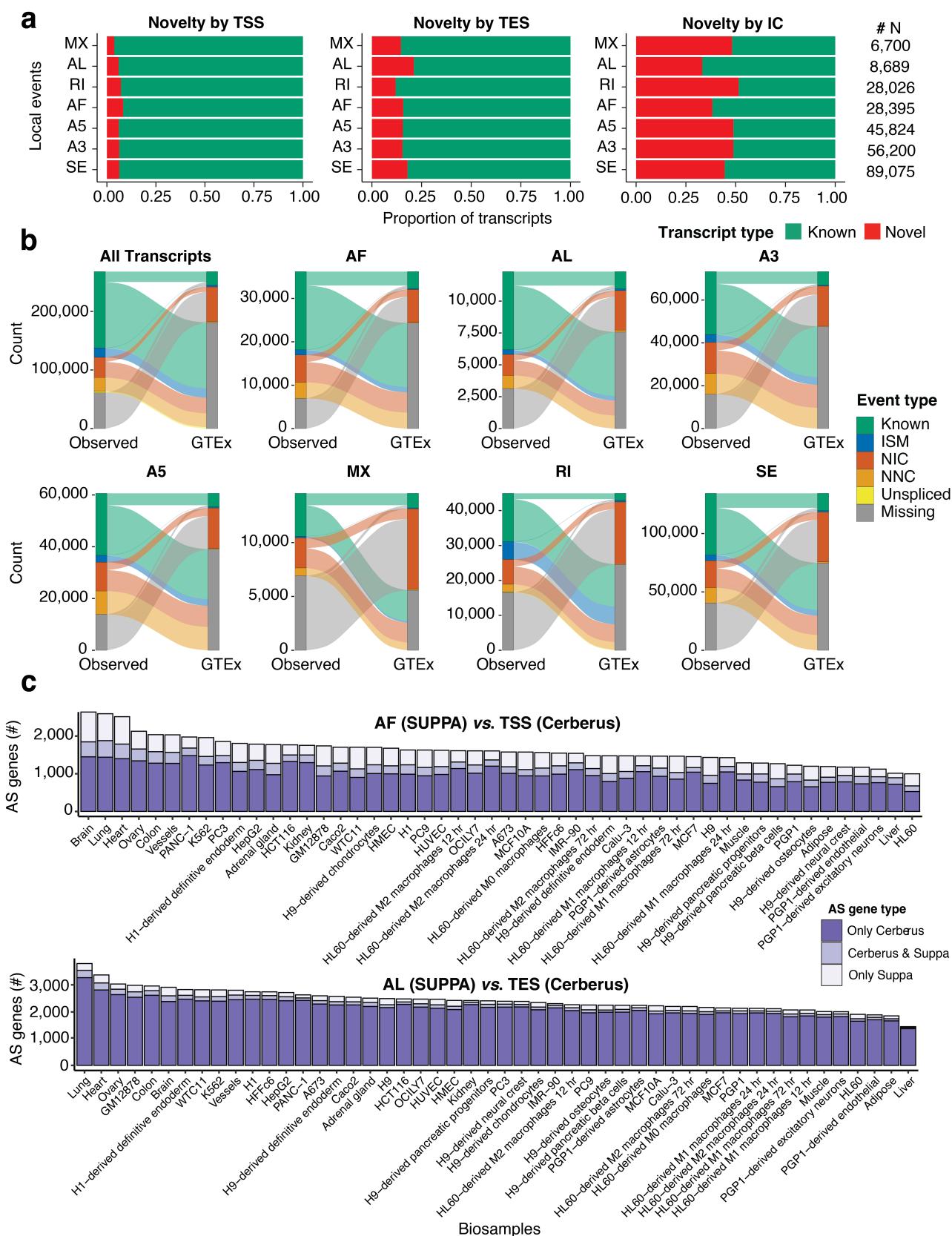


Figure S8. Alternative splicing (AS) detection by SUPPA and Cerberus. **a**, Barplots showing, for each type of local AS event detected by SUPPA (y axis), the proportion of observed known and novel transcripts identified by Cerberus (x axis), based on novel events at the TSS, TES, or EC. **b**, Sankey plots showing the number of transcripts classified as Known, Novel In Catalog (NIC), Novel Not in Catalog (NNC), Unspliced, or Missing by Cerberus and GTEex. In the first panel we show the numbers for all transcripts, while in the rest of the panels we focus on transcripts undergoing specific types of AS events. **c**, Barplot showing, for each LR-RNA-seq sample (x axis), the number of observed AS genes identified by both Cerberus and SUPPA (pink), only by Cerberus (dark gray), or only by SUPPA (light gray). In the upper panel we compare Alternative-First (AF)-AS genes by SUPPA and TSS-AS genes by Cerberus. In the lower panel we compare Alternative-Last (AL)-AS genes by SUPPA and TES-AS genes by Cerberus.

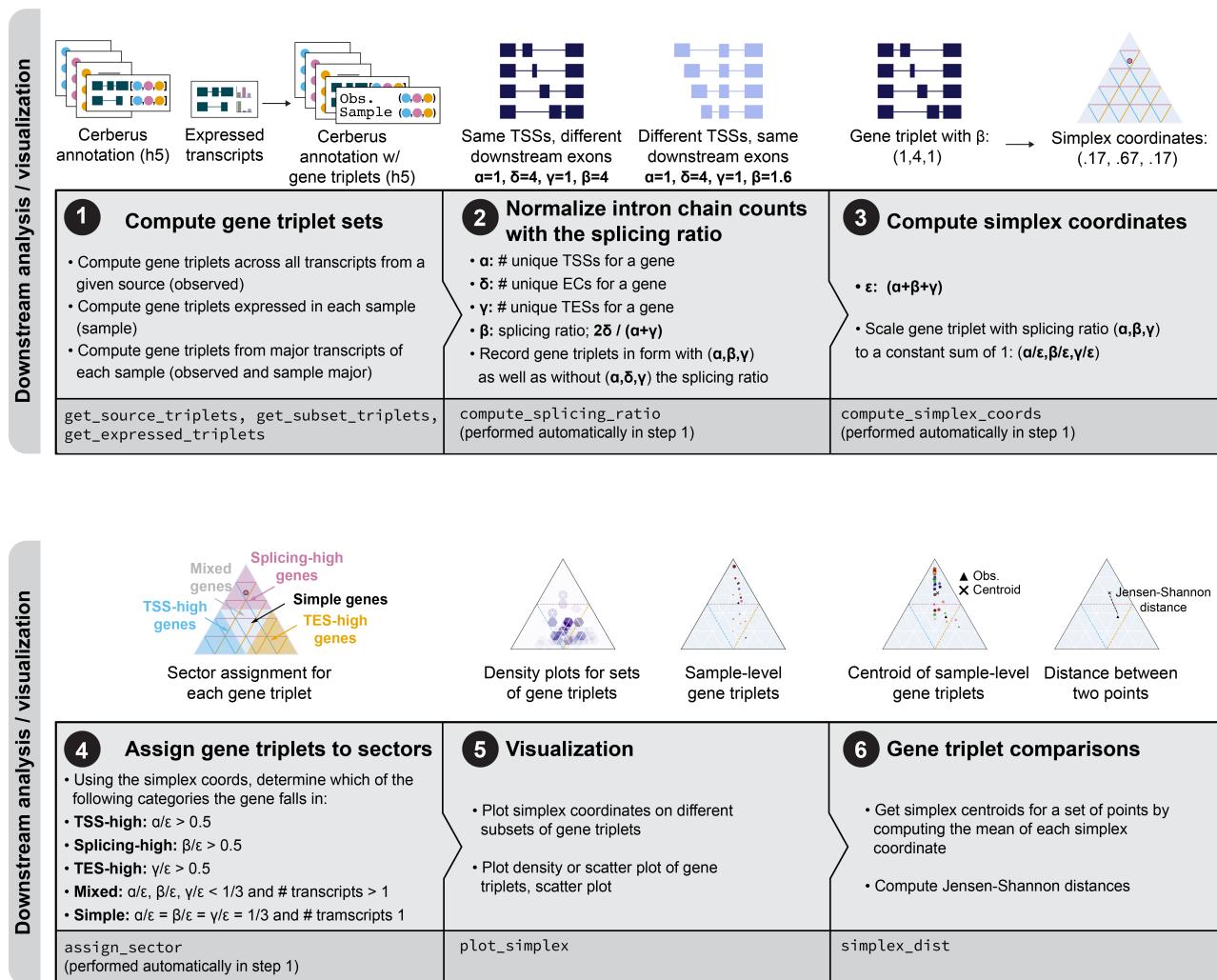
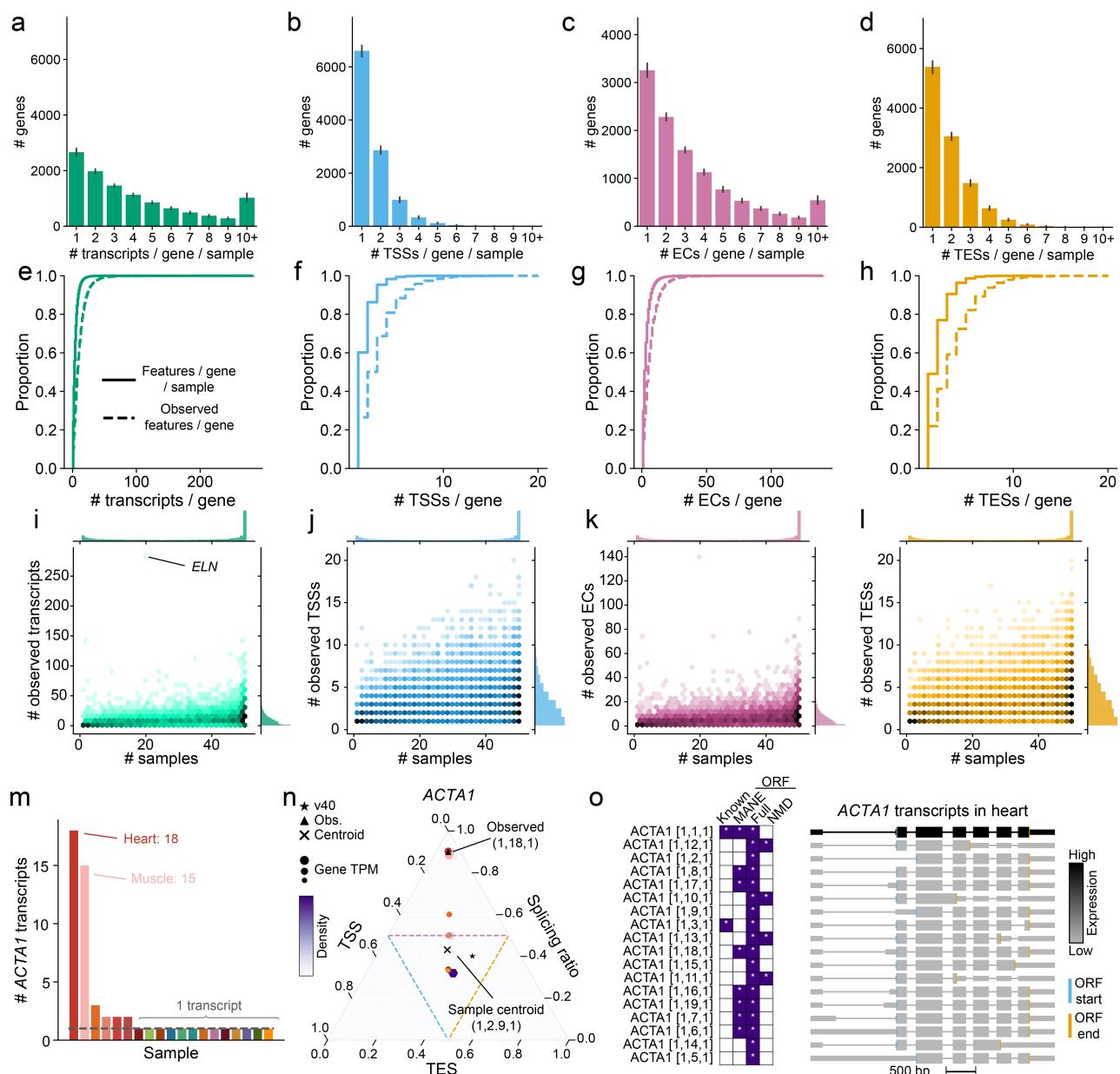


Figure S9. Overview of gene triplet based downstream analysis and visualization with Cerberus.



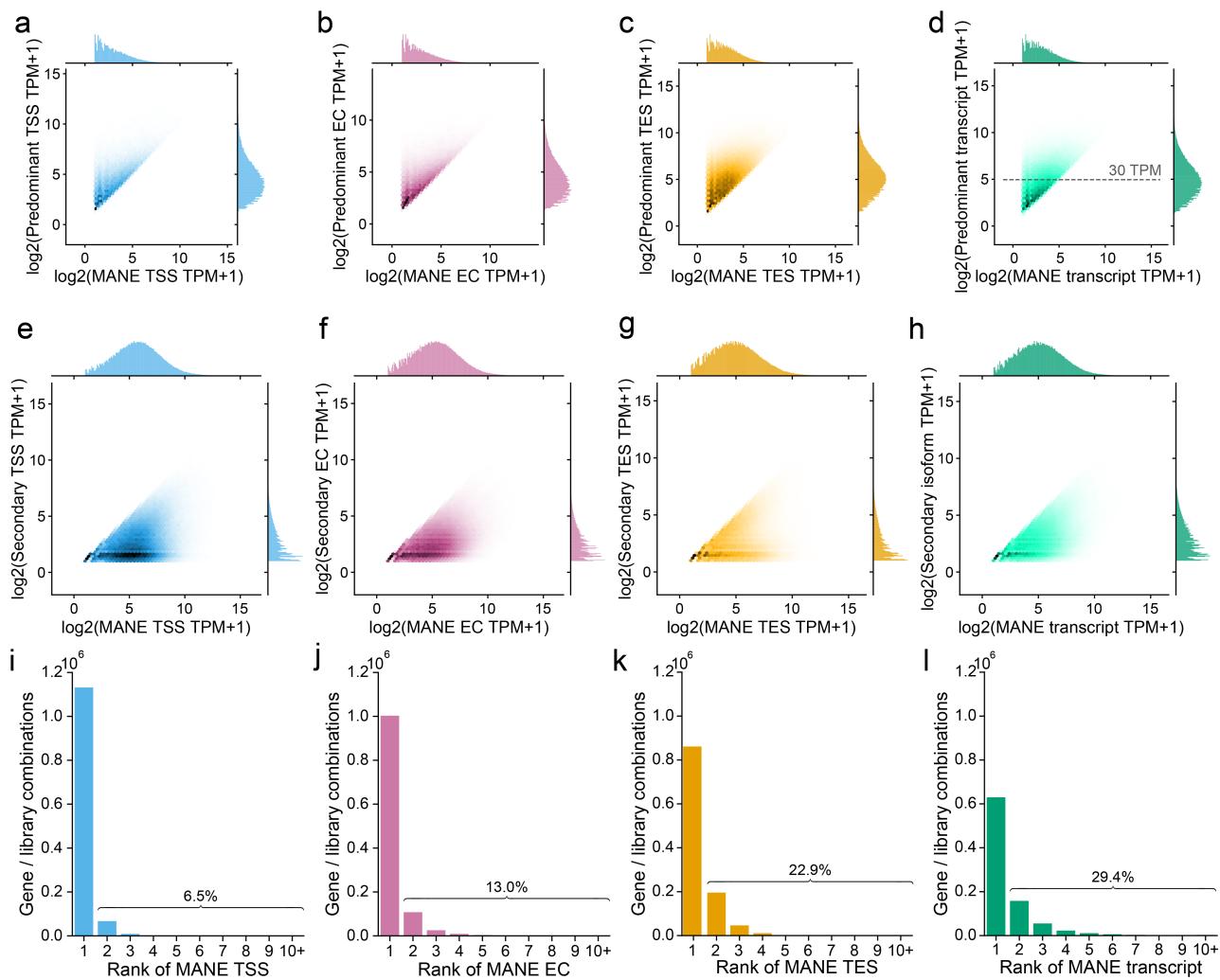


Figure S11. Rank and expression of predominant and MANE transcripts. **a-d**, For protein coding genes with MANE transcripts from GENCODE v40 where the predominant transcript or triplet feature is not the MANE transcript or triplet feature but is still expressed, expression of the predominant vs. the expression of the MANE **a**, TSS **b**, EC **c**, TES **d**, transcript. **e-h**, For genes where the MANE triplet feature or transcript is the predominant one and a secondary triplet feature or transcript is expressed, expression of the secondary vs. MANE **e**, TSS **f**, EC **g**, TES **h**, transcript. **i-l**, Rank of MANE **i** TSS **j**, EC **k**, TES **l**, transcript in each library where it is expressed.

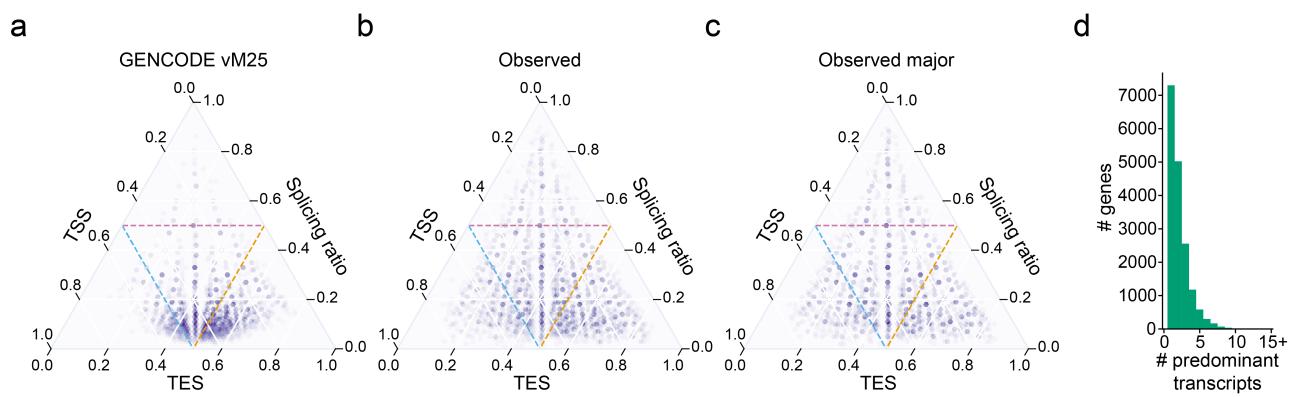


Figure S12. Rank and expression of predominant and MANE transcripts. **a-c**, Gene structure simplices for the transcripts from protein coding genes that are **a**, annotated in GENCODE vM25 where the parent gene is also detected in our mouse LR-RNA-seq dataset, **b**, the observed set of transcripts, those detected ≥ 1 TPM in the mouse ENCODE LR-RNA-seq dataset, **c**, the observed major set of transcripts, the union of major transcripts from each sample detected ≥ 1 TPM in the mouse ENCODE LR-RNA-seq dataset. **d**, Number of unique predominant transcripts detected ≥ 1 TPM across samples per gene.