Title

Characterizing transcript diversity using long-read RNA sequencing

Author

Reese, Fairlie

Publication Date

2023

UNIVERSITY OF CALIFORNIA,
IRVINE

Characterizing transcript diversity using long-read RNA sequencing

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Fairlie Reese

Dissertation Committee:
Ali Mortazavi, Chair
Lee Bardwell
Klemens Hertel

2023

DEDICATION

To Kiki and Junior

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my parents for nurturing my curiosity for the natural world, providing me with opportunities without which I surely would not be here today, and being supportive of me through the highs and lows.

I would also like to thank my Orange County friends that somehow managed to transcend my undergraduate years into graduate school. I have greatly appreciated our continued friendship. Of course I also have to thank the friends I made from UCI. Special thanks to Ria, who took me on an unforgettable trip to India; and my GenPALS co-leaders and regular participants whose perspectives from projects far different than my own have certainly influenced how I have approached my work.

From within the Mortazavi Lab, I of course want to thank everyone. I have truly learned something from each and every one of you. Special thanks to Dana and Gaby, my first project partners in the lab who helped orient me to the group, the projects, and UCI. Liz, my contemporary who started at the same time as I, for being an excellent project partner, a great person to talk about things wholly unrelated to science, a fantastic co-TA for COSMOS and D132, and the best cat auntie I could possibly ask for. Narges, for giving me computational-only solidarity in the lab, being a great bay-mate, and feeding me Persian food. Jaz, for your insightful questions and your willingness not only try new things but your drive to understand them deeply. Thanks also to the amazing team that currently keeps things running in the lab and are responsible for the firehose of data that I now have to reckon with: Heidi, Liz, Ghassan, Parvin; and in the past Cassie and Gaby.

Of course I also want to thank Ali for his mentorship and the amazing opportunities that I would not have had outside of his lab.

Thanks to my collaborators at Caltech, particularly Barbara, Brian, and Diane for their constant feedback and expertise.

I want to extend a general thanks to all of my ENCODE, LRGASP, and IGVF collaborators. Being in big consortia is equal parts glamorous and arduous but I have become a better scientist due to the insightful and opinionated discussions that I would have not had otherwise.

Last but certainly not least I would like to thank Sam, who is my at-home collaborator. Your love and insight have been invaluable. I am so thankful we have gone through graduate school together and I immensely look forward to continuing to grow together as scientists and individuals.

VITA

Fairlie Reese

EDUCATION

Doctor of Philosophy in Biological Sciences **2023**
University of California, Irvine *Irvine, CA*

Master of Science in Biological Sciences **2022**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Bioengineering: Bioinformatics **2017**
University of California, San Diego *San Diego, CA*

PUBLICATIONS

* These authors contributed equally

Published

1. Z Liu, G Quinones-Valdez, T Fu, M Choudhury, **F Reese**, A Mortazavi, X Xiao. L-GIREMI uncovers RNA editing sites in long-read RNA-seq. *Genome Biology.* (2023).
2. N Rezaie, **F Reese**, A Mortazavi. PyWGCNA: A Python package for weighted gene co-expression network analysis. *Bioinformatics* (2023).
3. S Morabito, **F Reese**, N Rahimzadeh, E Miyoshi, and V Swarup. High dimensional co-expression networks enable discovery of transcriptomic drivers in complex biological systems. *Cell Reports Methods.* (2023).
4. **F Reese***, E Rebboah*, K Williams, G Balderrama-Gutierrez, C McGill, D Trout, I Rodriguez, H Liang, BJ Wold, and Ali Mortazavi. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. *Genome Biology.* (2021).
5. A Krumpel, A Rice, KE Frasier, **F Reese**, JS Trickey, AE Simonis, JP Ryan, SM Wiggins, A Denzinger, H Schnitzler, and S Baumann-Pickering. Long-Term Patterns of Noise From Underwater Explosions and Their Relation to Fisheries in Southern California. *Frontiers Marine Science.* (2021).
6. JE Moore, X Zhang, SI Elhajjajy, K Fan, HE Pratt, **F Reese**, A Mortazavi, and Z Weng. Integration of high-resolution promoter profiling assays reveals novel, cell type-specific transcription start sites across 115 human cell and tissue types. *Genome Research.* (2021).

7. **F Reese**, and A Mortazavi. Swan: a library for the analysis and visualization of long-read transcriptomes. *Bioinformatics.* (2020).

8. M Movassat, E Forouzmand, **F Reese**, KJ Hertel. Exon size and sequence conservation improves identification of splice-altering nucleotides. *RNA.* (2019).

In review / preparation

1. E Miyoshi*, S Morabito*, CM Henningfield, N Rahimzadeh, SK Shabestari, S Das, N Michael, **F Reese**, Z Shi, Z Cao, V Scarfone, MA Arreola, J Lu, S Wright, J Silva, K Leavy, IT Lott, E Doran, WH Yong, S Shahin, M Perez-Rosendahl, E Head, K Green, and V Swarup. Spatial and single-cell transcriptomic analysis of genetic and sporadic forms of Alzheimer's disease. *bioRχiv* (2023).

2. **F Reese**, BA Williams, G Balderrama-Gutierrez, D Wyman, MH Çelik, E Rebboah, N Rezaie, D Trout, M Razavi-Mohseni, Y Jiang, B Borsari, S Morabito, H Liang, C McGill, S Rahmanian, J Sakr, S Jiang, W Zeng, K Carvalho, A Weimer, LA Dionne, A McShane, K Bedi, S Elhajjajy, J Jou, I Youngworth, I Gabdank, P Sud, O Jolanki, JS Strattan, M Kagda, MP Snyder, BC Hitz, JE Moore, Z Weng, D Bennet, L Reinholdt, M Ljungman, MA Beer, MB Gerstein, L Pachter, R Guigó, BJ Wold, A Mortazavi. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *In revision.* 2023

3. BC Hitz, JW Lee, O Jolanki, KS Meenakshi, K Graham, P Sud, I Gabdank, JS Strattan, CA Sloan, T Dreszer, LD Rowe, NR Podduturi, VS Malladi, ET Chan, JM Davidson, M Ho, S Miyasato, M Simison, F Tanaka, Y Luo, I Whaling, EU Hong, BT Lee, R Sandstrom, E Rynes, J Nelson, A Nishida A Ingersoll, M Buckley, M Frerker, DS Kim, N Boley, DE Trout, A Dobin, S Rahmanian, D Wyman, G Balderrama-Gutierrez, **F Reese,** NC Durang, O Dudchenko, D Weisz, SSP Rao, A Blackburn, D Gkountaroulis, S Mahdi, M Olshanky, Y Eliaz, D Nguyen, I Bochkov, MS Shamim, R Saad, A Erez, T Gingeras, S Heath, M Hirst, WJ Hirst, A Kundaje, A Mortazavi, BJ Wold, JM Cherry. The ENCODE Uniform Analysis Pipelines. *bioRχiv.* (2023).

4. HC Happ, PN Schneider, JH Hong, E Goes, M Bandouil, CG Biar, A Ramamurthy, **F Reese**, K Engel, S Weckhuysen, IE Scheffer, HC Mefford, JD Calhoun, GL Carvill. Long-read sequencing and profiling of RNA-binding proteins reveals the pathogenic mechanism of aberrant splicing of an *SCN1A* poison exon in epilepsy. *bioRχiv.* (2023).

5. JE Childs, S Morabito, S Das, C Santelli, V Pham, K Kusche, V Alizo Vera, RR Campbell, DP Matheos, **F Reese**, A Mortazavi, V Swarup, and MA Wood. Medial Habenula *Nr4a2* is necessary for reinstatement of cocaine self-administration and related transcriptome changes identified using single nuclei RNA-seq. *In review.* (2022).

6. **F Reese***, F Pardo-Palacios*, S Carbonell-Sala*, M Diekhans*, C Liang*, D Wang*, B Williams*, M Adams, A Behera, J Lagarde, H Li, A Prjibelski, G Balderrama-Gutierrez, MH Çelik, M De María, N Denslow, N Garcia-Reyero, S Goetz, M Hunter,

J Loveland, C Menor, D Moraga, J Mudge, H Takahashi, A Tang, I Youngworth, P Carninci, R Guigó, H Tilgner, BJ Wold, C Vollmers, G Sheynkman, A Frankish, KF Au, A Conesa, A Mortazavi, and A Brooks. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *In review.* (2021)

7. **F Reese***, D Wyman*, G Balderrama-Gutierrez*, S Jiang, S Rahmanian, S Forner, D Matheos, W Zeng, B Williams, D Trout, W England, S Chu, RC Spitale, AJ Tenner, BJ Wold, and A Mortazavi. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRχiv.* (2020).

SOFTWARE

Swan `https://github.com/mortazavilab/swan_vis`
Python-based analysis and visualization of full-length transcriptomes.

LR-Splitpipe `https://github.com/fairliereese/LR-splitpipe`
Demultiplexing of long-read single-cell data barcoded with Parse Biosciences.

Cerberus `https://github.com/fairliereese/cerberus`
Triplet based harmonization and transcript diversity analysis of full-length transcriptomes.

SELECTED PRESENTATIONS

Mid Atlantic Splicing Conference **2023**
Keynote speaker

American Society for Human Genetics **2022**
Accepted speaker

Intelligent Systems for Molecular Biology, iRNA **2022**
Accepted speaker

ENCODE Consortium Meeting **2022**
Invited speaker

ENCODE Consortium Meeting **2021**
Invited speaker

Intelligent Systems for Molecular Biology, iRNA **2020**
Accepted speaker

ENCODE Long-read RNA-seq Meeting **2019**
Invited speaker

AWARDS

Developmental and Cell Biology Research Excellence **2022**

LEADERSHIP

UCI GenPALS Co-founder **2020-Present**

WORKSHOPS

European Conference on Computational Biology LRGASP Workshop **2022**

UCI GenPALS Intro to scRNA-seq Workshop **2021, 2022, 2023**

UC Davis IsoSeq Workshop **2021**

TEACHING EXPERIENCE

Teaching Assistant **2021**
UCI COSMOS (Genes, Genomes, and Skeletal Muscle Dystrophies) *Irvine, CA*

Teaching Assistant **2020, 2021**
Intro to Precision Medicine (D132) *Irvine, CA*

ABSTRACT OF THE DISSERTATION

Characterizing transcript diversity using long-read RNA sequencing

By

Fairlie Reese

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2023

Ali Mortazavi, Chair

Alternative transcripts arise from the same gene via alternative TSS usage, splicing, and polyA site choice. Such transcripts can give rise to functional disparities in protein structure, post-transcriptional regulation, and translational efficiency. Moreover, their expression in appropriate spatiotemporal contexts is a key feature of eukaryotic genomes. However, detecting and quantifying these transcript isoforms across tissues, cell types, and species has been challenging due to their longer lengths compared to the short reads typical of standard RNA-seq. In contrast, long-read RNA-seq (LR-RNA-seq) provides complete transcript structures, enabling investigation of transcript features and usage with greater fidelity.

Here, I describe my work on application of LR-RNA-seq to characterizing and comparing full-length transcriptomes. First, I describe Swan, a software library I developed to facilitate visualization of full-length transcripts and to compare transcript usage between biological conditions. Next, I describe the ENCODE4 human and mouse LR-RNA-seq datasets, where I applied a novel triplet-based framework to harmonize and classify transcripts that share transcript start sites, exon junction chains, and transcript end sites. Lastly, I discuss the application of our single-nucleus LR-RNA-seq technique (LR-Split-seq) on two genetically distinct mouse strains to uncover cell type and genotype-specific transcript usage patterns. Collectively, these projects form a solid foundation for future analyses of long read transcrip-

tomes to quantify changes in transcript diversity and transcript usage between samples, cell types, and genotypes within and between species.

Chapter 1

Introduction

1.1 Abstract

Alternative transcript isoforms can arise from a single gene that can result in functional differences in protein structure or function, post-transcriptional regulation, and translational efficiency. The balance and specificity of transcript isoform expression play crucial roles in vital biological processes such as development and differentiation, whereas disruptions in appropriate transcript isoform expression can lead to disease. While short-read RNA-seq has been widely used to study gene expression over the past decade, but lacks the capability to profile full-length transcript isoforms, which complicates the identification of both baseline and disrupted transcript expression profiles. In contrast, long-read RNA-sequencing (LR-RNA-seq) platforms, though lower in throughput, offer the advantage of sequencing full-length transcript isoforms. Here, I describe the fundamentals of gene and transcript expression and discuss various technologies, including LR-RNA-seq, used to study transcription. I discuss prior findings based on these assays that attempt to characterize transcript diversity and full-length transcriptomes. Finally, I explore how alternative transcript iso-

forms have been profiled in single cells or nuclei using both long and short-read RNA-seq methods.

1.2 Introduction

The human genome, gene expression, and transcript isoforms

With few exceptions, the genomic sequence in each cell of a multicellular organism is the same. Specialized functions of distinct organs or cell types are driven by distinct programs of gene expression that activate or repress genes necessary for the cell to perform its prescribed function. Before the human genome was first sequenced, it was estimated that humans, as self-proclaimed highly complex organisms, had around 100,000 genes[1]. However, after the human genome was sequenced and gene prediction efforts catalogued the transcribed regions of the human genome, they were surprised to identify only around 25,000 protein coding genes[2,3]. Given that seemingly less complex organisms such as *C. elegans* have a similar number of protein coding genes as humans[4], it follows that there is a different mechanism that gives rise to the apparent complexity of higher eukaryotes. One mechanism by which this complexity is derived is through alternative transcripts from the same gene encoded for in DNA[5]. When RNA is transcribed from a gene, the mature mRNA products can take multiple forms called transcript isoforms that are influenced by promoter choice, alternative splicing, and by polyA site choice (Fig. 1.1a)[6–9]. The subsequent changes introduced by these options can give rise to different RNA transcripts and subsequent protein products that are distinct in their functional roles. In this way, with only 25,000 protein coding genes, humans can hypothetically produce an order of magnitude more distinct proteins that characterize cellular identity and distinct transcriptional programs from one another. Alternative transcripts can additionally be subject to different post-transcriptional fates, which are in part governed by the content of their 5' and 3' untranslated regions (UTRs).

The 3' UTR often contains microRNA or RNA binding protein (RBP) binding sites which can either enhance or reduce the potential for a protein coding mRNA to be translated[6,7]. Regulatory elements can also exist in the 5' UTR, such as microORFs, which can additionally affect translational efficiency[10]. Both UTRs have also been found in some cases to harbor sequence-based localization signals that affect where the transcript localizes in the cell[11]. To holistically assess transcriptional diversity across biological contexts, it is crucial to consider differences in splicing, as well as the 5' and 3' ends, due to their significant influence on the functional outcome of transcribed RNA products.

The functional roles of alternative transcript isoforms

There are many examples of differential function of protein products that result from transcript choices. Transcripts of the same gene can also have strikingly opposing effects. In one case, the gene *BCL2L1*, can produce two transcripts, one of which is pro-apopototic and the other which is anti-apoptotic[12]. Alternative transcript usage by means of splice alteration has been estimated to account for 15-60% of known genetic risk disease variants[13,14]. For instance, Duchenne's muscular dystrophy is caused by missplicing of the dystrophin gene which introduces a premature stop codon[15]. Polygenic diseases are also often influenced by alternative transcript usage. Disruptions in the relative abundances of different *MAPT* transcripts have been implicated in neurodegenerative diseases such as Alzheimer's Disease[16].

Many biological processes rely on the expression of specific transcript isoforms of a gene in the appropriate spatiotemporal context. Neurodevelopmental processes such as neural stem cell differentiation, neuronal migration, and synaptogenesis rely on specific transcript isoforms and splicing factors, which consequently affect the resultant transcripts[17]. There are substantial changes in alternative splicing during muscle development and repair[18,19]. Given that the regulatory potential and function of transcripts can differ drastically from one another, transcript isoforms are important to study because they are essential for defining the identity of a biological context.

Technological limitations to studying transcript identity and expression

Transcript isoforms and gene expression have been studied using many different assays in the past. Expressed sequence tags (ESTs) relied on cloning of cDNA and subsequent sequencing of each clone[20]. While they are useful for the discovery of transcribed genomic regions, sequencing was typically single-pass and done using capillary sequencing, meaning both that ESTs are difficult to use for estimating gene expression levels and that they are not long enough to yield full-length transcript isoform information, as the average length of protein coding transcripts from GENCODE v40 is 1.7 kb[21]. Separately, microarrays, which contain arrays of specific complementary oligonucleotide probe sequences from transcribed genes, were used to profile gene expression[22]. As microarrays use predesigned probes for target capture, they require *a priori* knowledge of genes and cannot profile novel transcribed regions in an unbiased manner. Furthermore, they can only capture specific transcript isoforms on the basis of the probe design, meaning that certain transcripts will be missed. The development of RNA-seq decoupled gene and transcript identification and quantification from probes for a more unbiased method[23]. However, when performed on short-read sequencing platforms such as Illumina, which have been the gold standard for years, RNA-seq reads are not long enough to capture the entire structure of a transcript without assembly and therefore cannot fully resolve each transcript isoform (Fig. 1.1b)[5].

Studying transcript isoforms using long-read RNA-seq

Long-read RNA-seq (LR-RNA-seq) can sequence reads that are long enough to capture the entirety of each transcript isoform (Fig. 1.1c). The two main long-read sequencing platforms are from Pacific Biosciences (PacBio) and Oxford Nanopore (ONT)[24–26]. PacBio sequencing works similarly to short-read Illumina sequencing in that cDNA molecules are sequenced based on fluorescence of bases added to a complementary DNA sequence during second strand synthesis[25,26]. By contrast, ONT flowcells sequence cDNA or unamplified RNA by pulling a single-stranded molecule through an engineered pore with a motor protein. Each base of the read is called by measuring the voltage change across the membrane where the pore sits, as different nucleotides will introduce different voltage changes[24].

Both technologies, which were plagued by high error rates and low throughput in their earlier years, have matured substantially. PacBio produces circular consensus (CCS) reads, which help improve the basecalling error rate. Rather than reducing the native error rate, the molecule of interest is sequenced repeatedly in a loop. This allows for multiple sequencing passes to be performed on the same insert fragment, which can be harnessed bioinformatically to create a consensus read with a current reported accuracy of 99.9%[27]. A similar circular consensus correction approach has been applied to ONT as well, albeit using a custom protocol[28]. ONT has been developing duplexed read technology such that the forward and reverse copies of a target sequence are read in succession, which provides more information about each base and allows for a decreased basecalling error rate. Currently, ONT's highest reported accuracy rate, using double-passed duplex reads and the most stringent basecalling settings, is 99.9%[29].

The throughput for long-read sequencing platforms has also improved dramatically. Recently, PacBio has commercialized a library preparation technique called MAS-Iso-seq, which optimizes the information yield of each run by concatenating multiple transcripts, which are on average around 1 kb long. Because the PacBio platform is often used to sequence genomic DNA where the fragments are often much longer, the optimal library size to yield the recommended number of CCS passes is 15-20 kb long. Thus, by concatenating multiple shorter constructs together to reach this target length, the number of reads from an LR-RNA-seq run has increased substantially[30]. Hypothetical maximum throughput for one PacBio flowcell using their newest sequencing platform (the Revio) and MAS-Iso-seq has now increased to around 100 million reads. The highest-throughput flowcell that ONT currently makes, the PromethION, has a hypothetical yield of 290 Gb of sequence[31]. With these recent advances in throughput and accuracy from both technologies, the field is now more than ever poised to answer key questions regarding both transcript identity as well as transcript expression.

Computational approaches to studying long-read RNA-seq

As the sequencing approaches for LR-RNA-seq mature, so too do the computational tools and workflows used to study such data. At the time of writing, an online database of software tools to process long read sequencing data has around 800 unique tool entries, which has increase from the 200 listed by the database in 2018[32]. For LR-RNA-seq specifically, tools roughly stratify by whether they are related to data preprocessing or downstream analysis.

Aside from platform-specific preprocessing steps such as CCS and basecalling, a typical LR-RNA-seq preprocessing includes the following steps:

- Read mapping, which is typically done using minimap2[33]
- Read annotation, which involves annotating each long read to its transcript of origin either with or without a reference transcriptome annotation
- Transcript quantification, which assigns each transcript an expression value

The task of read annotation has been particularly challenging for researchers. This is often done by either clustering reads that share the same sequence of splice junctions (also referred to as intron chain or exon junction chain) or by determining which reference transcript from an annotation such as GENCODE[21] shares the same exon junction chain[34–37]. While simple in theory, concordance of exon junction chain is complicated by small insertions or deletions, called microindels, which are a well-known artifact of both the ONT and PacBio platforms[25]. These introduce small differences at the 5' and 3' ends of splice junctions which can cause algorithms that rely on exact exon junction chain matches to categorize reads as from distinct transcripts even if they are simply technical artifacts. To this end, some preprocessing workflows include correction of such artifacts[35,38].

Another complication arises from incomplete reads. Though both the ONT and PacBio platforms have the capacity to sequence reads long enough to capture the majority of transcripts, there often are many instances of what appear to be "incomplete" transcripts that share a portion of the exon junction chain of either a reference transcript or another transcript already seen in the data. These likely result from a) degraded RNA in the input sample, b) incomplete sequencing of an existing molecule (which is a more a more noticeable problem the longer the transcript gets, such as transcripts from the *TTN* gene[39]), or c) from true biological instances of shorter versions of transcripts. It is difficult to distinguish true instances of alternative transcript usage via shortening from the experimental artifacts of degradation or incomplete sequencing, and different workflows approach this issue in various ways. Some rely on an updated version of transcript assembly that is sensitive to the nature of LR-RNA-seq[40]. Others take a peak calling or clustering approach of read starts and ends to distinguish artifacts from the biological truth via reproducibility[37,41]. Many other tools rely on colocalization of external validation from 5' and 3' specific assays, such as CAGE and PAS-seq, to determine the true transcript ends[37,42]. Additional tools incorporate information about whether or not the polyA motif is found where a read's 3' end is[41,43].

Quantification from LR-RNA-seq poses an additional challenge for the field. In theory, each read, whether generated using PacBio or ONT, should represent precisely one molecule, simplifying the quantification task. Many tools do take this simple route of counting reads based on which transcript they were annotated to[35,36]. However, in practice, quantification is intrinsically and sensitively tied to transcript annotation[44], and some tools take common structural features of the identified transcripts into account and thus can assign one read count across multiple transcripts[43]. Other approaches harness statistical models that incorporate information about the each read, including its quality and likeliness to be full-length, for quantification[45,46].

Current analytical tools downstream of generating transcript annotations focus on transcript characterization, transcript isoform switching tests, and transcript visualization. Transcripts are often first characterized by their novelty with respect to a reference annotation, though as previously mentioned some annotation strategies are reference-free[47,48]. This is a step that does not really share an analogous short-read RNA-seq workflow step, as one of the unique advantages of LR-RNA-seq is its potential to discover unannotated full-length transcripts. Transcript novelty categories as coined by SQANTI[42] are frequently used to describe the relationship of an identified transcript to an existing reference annotation (Fig. 1.2). Full splice match (FSM, or known) transcripts share the entirety of their exon junction chain with an existing transcript, incomplete splice match (ISM) transcripts share a subsequence of their exon junction chain with an existing transcript, novel in catalog transcripts (NIC) have all their splice sites in the reference but at least one unannotated splice junction, and novel not in catalog (NNC) transcripts have at least one unannotated splice site. Additionally, transcripts can be intergenic, or lie outside annotated gene boundaries; or genomic, which are monoexonic, often intron-containing transcripts that fall within annotated gene boundaries[42]. Further subclassifications of different novelty categories have been developed as well by other groups[37,46].

Characterization of transcripts can be done via functional prediction, which often relies on defining a reference transcript for a gene to predict the change in function that a transcript might have with respect to the reference. Defining a reference transcript is a rather problematic and circular issue, however, and suffers from its lack of sensitvitiy to what might be the most highly-expressed transcript or transcripts from a gene in a given biological context[49,50]. Functional prediction tools can predict changes at the 5' and 3' end as well as the in the protein coding sequence. Protein coding sequence predictions are often followed up by assessment for gain or loss of protein domains[51–53].

Finding genes where transcripts are differentially abundant or expressed across distinct biological contexts is a unique problem. Classic differential expression tests that are often applied on the gene level can easily be applied on the transcript level as well, but are confounded by differences solely driven by changes in gene expression[54]. To combat this, researchers often use isoform switching tests instead, which are also referred to as differential isoform expression or differential isoform usage tests[52,54–58]. These differ from basic differential expression tools as transcript expression is analyzed independently of the expression of the gene that it comes for; represented as the percentage of the gene's expression that is attributable to that specific transcript. This metric, analogous in spirit to the percent-spliced-in (Ψ) metric often used to quantify relative occurrence of individual AS events, is referred to as percent isoform (π)[57,59]. These tests are more sensitive to phenomena such as changes in the most highly-expressed transcript across biological contexts and can yield insights into sample specificity of alternative transcripts.

Finally, visualization of alternative transcripts is an active field of LR-RNA-seq software development. Some alternative events are difficult to see at genomic scale using traditional genome browser style visualizations, and even if they are easy to see, they can still be difficult to differentiate by eye for complex transcripts. Furthermore, genome browser tools offer no solutions to visualizing transcript expression values across multiple samples. There have been many tools developed to address these issues, which facilitate the thorough exploration of LR-RNA-seq data. Some tools provide visualization options for transcript structure outside of the genome browser setting[56,60,61]. Other tools display percent isoform expression values for different transcripts from the same gene to enable visual understanding of transcript isoform switches[52,55,56]. Additional tools integrate visual tools with functional prediction tools and display important features such as protein domain content alongside plotted transcript models[52].

The impact of library preparation, sequencing platform, and software tools on long-read RNA-seq analysis tasks

The performance of different software tools, library preparation techniques, and sequencing platforms on the problems of both transcript annotation and quantification has recently been assessed by the LRGASP consortium[44]. This effort sequenced samples using different library preparation techniques, such as direct RNA, CapTrap cDNA[62], R2C2 cDNA[28], and standard oligo dT primed cDNA; and different sequencing platforms (ONT and PacBio). Higher read quality and longer read length of the sequencing platform were found to be associated with improved performance on the transcript annotation task, and they therefore recommended to use standard oligo dT primed cDNA libraries on the PacBio sequencing platform. However, with the latest improvements in ONT error rate, it is unclear whether this recommendation would hold. For the purposes of transcript quantification, they recommend cDNA sequenced on the ONT platform; likely due, at the the time, to ONT's increased throughput over PacBio. Just as the library preparation method and sequencing platform affect the results, so too do the software tools used; and the recommendations for which approaches to use vary depending on whether the goal is quantification or transcript annotation. Therefore, it is clear that careful considerations must be taken when designing experiments and analytical pipelines when working with LR-RNA-seq data.

Prior work on studying transcript diversity and alternative splicing

It is an important task to catalog and to characterize transcripts based on the biological contexts they are in to understand how they might be perturbed upon genetic variation, disease state, or during development. Many studies have looked at the complexity of the eukaryotic transcriptome with respect to alternative splicing (AS) and transcriptional diversity. Generally, when the methodology used to study transcript diversity is length limited, such as in the case of microarrays and short-read RNA-seq, this diversity is quantified on the level of exon-level AS events. A study using microarray HapMap data in 2007 found that individual AS events are more variable between human individuals than gene expression was, which indicates the role of genetic diversity in driving choice of transcript isoform[63]. This result has been replicated by other groups, including those done using short-read RNA-seq data rather than microarrays[64]. Several studies have attempted to assess the fidelity or stochasticity of the splicing machinery by relating the relative gene structure in terms of exon and intron content and length to the number of observed transcript isoforms or AS events. Generally, they found that genes with more and longer introns have more splicing "noise" and therefore yield a greater number of transcript isoforms for these genes[65–67]. However, these studies claim opposing effects on the number of AS events or transcript isoforms that higher gene expression levels yield[65,66]. Unsurprisingly, both studies suffer from their inability to assess full-length content of each transcript (using ESTs and short-read RNA-seq) as well as from their limited pool of samples. Indeed, it has also been shown that many characterized AS events are tissue-specific, which suggests the increased importance in transcript specificity as a determinant of tissue or cell type identity, and that key features of transcript diversity will inevitably be missed when not considering a variety of samples[67]. It has also been shown that splicing diversity is higher in humans than in mouse[68] and has been suggested that the increased splicing diversity in humans could be due to the altogether longer intron sizes in human and association of intron size and splicing rate, as well as the tendency of primate-specific Alu elements to exonize[66,69,70]. In 2013, it was reported that across 16 tissues and 5 cell lines that, despite substantial diversity amongst detected transcripts, the most ubiquitously-expressed genes typically use the same transcript as the most highly-expressed one across their samples[71].

The Encyclopedia of DNA elements (ENCODE) consortium has been a mainstay in genomics since its pilot in 2003. Broadly, the goal of this consortium is to catalog and characterize all the functional elements in the genome[72]. While a large part of the consortium's efforts have been to characterize the state of the DNA in terms of epigenetic marks, chromatin accessibility, and 3D conformation, another critical component to understanding the regulatory impact of the genome is by characterizing the actual transcriptional output of the genome. To this end, ENCODE has always had an additional interest in high-throughput gene expression assays, whether it be from microarrays or RNA-seq. In 2012 as a part of ENCODE2, they demonstrated that 75% of the human genome is capable of being transcribed. Additionally, they studied splicing using short-read RNA-seq as many other groups have. Using 15 cell lines, they reported the capacity of the same cell lines to express multiple transcript isoforms of the same gene, notably present in many cases at varying levels of expression[73]. For ENCODE3 in 2020, they showed that while alternative splicing is a primary contributor to cell type identity, it does not contribute as highly to the overall transcriptional programs for major cell types[74]. However, each of these studies was performed using short-read RNA-seq. For the final phase of ENCODE, our group was funded to produce LR-RNA-seq data on ENCODE samples in addition to the prior transcriptional assays we have performed in the past.

The first big consortium-level analysis of full-length transcriptomes from LR-RNA-seq came out from GTEx in 2022 with ONT cDNA sequencing of 88 human tissue and cell line samples that identified around 70,000 novel transcripts. Stratified by type of AS event, they found that intron retention events and mutually exclusive exons were the least likely to already have been annotated, which is unsurprising given that these events can span transcript regions that might be too long to capture with short read data. As they also had access to sample genotypes, they focused on the impact of alternative alleles on gene expression and transcript usage. They found that when alternative alleles drove changes in transcript usage, they same genes also frequently were differentially expressed. Additionally, they performed variant effect prediction using the context of their novel transcripts and found that variants with uncertain significance were the ones most often reclassified, indicating that unannotated transcripts might help explain the linkage of uninterpretable variants to pathogenicity[75].

The coordination between events that lead to transcript diversity has also been investigated. If each AS event, TSS selection, and TES selection were completely independent from one another (barring events that are incompatible), one would expect to see orders of magnitude more transcripts resulting from any multiexonic gene. However, it is well known that despite some genes showing stochastic noise in alternative splicing decisions[65], many pairwise events are readily coordinated. Exon pairs can be mutually inclusive or exclusive[76], splicing efficiency of different introns has been linked to 3' end cleavage[77], and more recently, it was shown that TSS selection influences both AS events and 3' end site choice[78]. Notably, it was demonstrated that genes with alternative TSSs have significantly more alternative TESs as well, suggesting that highly transcriptionally diverse genes differ on multiple axes of variability[78].

Single-cell long-read RNA-seq

Single-cell and single-nucleus RNA-seq is a relatively young technology that enables researchers to profile gene expression in individual cells. This is achieved by assigning a unique sequence-based barcode to each cell, thus facilitating its identification. Currently, two widely used approaches for cell barcoding are microfluidics-based barcoding, where each cell is encapsulated in an oil droplet with its barcode; and combinatorial barcoding, which involves ligating a sequence of unique barcodes to determine cell identity[79–82]. These methods allow the identification of genes in specific cell populations that might otherwise be obscured within the complex mixture of cell types present in a tissue sample. In bulk samples, gene expression signals from rare or underrepresented cell populations often get overwhelmed by more abundant cell types[83]. Single-cell RNA-seq is now routinely used to characterize a multitude of biological phenomena where changes in gene expression at a cell type or cell state resolved resolution can yield additional insights[84,85].

For as long as single-cell RNA-seq has been around, groups have been using it to study alternative splicing[86–88]. However, the traditional approaches that involve splice junction counting to quantify AS events suffer immensely from the sparsity of single-cell data, as only the fraction of reads that span exon-exon junctions are informative on top of the already sparse single-cell data[79]. Furthermore, the very popular 10x single-cell platform only uses

3' end RNA selection, which limits the number of unique exon-exon junctions possible to observe. This sparsity has led to the false appearance of bimodal splicing, where an AS event is either always or never used, in each individual cell[89,90].

In recent years, as throughput for LR-RNA-seq has improved, multiple groups have started to perform single cell and single nucleus LR-RNA-seq experiments. As with short-read sequencing, these have been done using both microfluidics and combinatorial barcoding approaches[91]. Such techniques have been readily applied to study the cell type specificity of alternative isoforms during development and in disease. For instance, it has been shown that individual cell types contribute more to observed heterogeneity in alternative splicing between tissues as compared to the aggregate of all cell types[57]. Single-cell LR-RNA-seq of cancer samples has shown that cancerous cell types express more transcript isoforms at lower levels, hinting at the cancer-specific dysregulation of transcriptional control[92]. Additionally, research has demonstrated that cell types that do not dominate their resident tissue exhibit a higher abundance of novel transcript isoforms in comparison to more highly-abundant cell types[93]. The cell type specificity of exon pair coordination has also been investigated, both for adjacent and distal exon pairs, which would be impossible to study using traditional short-read RNA-seq[76]. Furthermore, the degree to which sub cell type, developmental age, or tissue region exhibits more control over the variability of transcript isoforms has been shown to vary drastically for different brain-resident cell types[94].

In my previous work, in conjunction with my co-first author Elisabeth Rebboah, we developed LR-Split-seq in 2021, a single-cell LR-RNA-seq approaches that uses Parse Bioscience's combinatorial barcoding strategy with PacBio LR-RNA-seq. We developed experimental and computational workflows needed to generate and process the data[95]. We applied this system to the C2C12 cell line model of myogenesis (muscle differentiation) and found that we were able to detect transcript isoform and TSS switches between distinct cell populations. Notably, we showed that the TSS switches were concomitant with chromatin opening using corresponding single-nucleus ATAC-seq data[96].

Of course, single-cell and single-nucleus LR-RNA-seq approaches are also faced with the same challenge of sparsity that comes with studying gene expression or AS events in single cells. A natural question given the sparsity is to what extent are the findings from existing single-cell LR-RNA-seq efforts limited by the difficulty in sampling. Single-cell LR-RNA-seq also comes with a unique set of challenges as compared to short-read approaches, some of the most daunting of which are related to the efficiency of capturing full-length transcripts in single cells or nuclei. Single-nucleus approaches, which are better suited for tissues that are difficult to dissociate or are multinucleated[97], definitionally enrich for the unprocessed RNA that is only present in the nucleus. This leads to the off-target capture of transcripts that have not been spliced or are in the process of splicing, which are almost entirely uninformative for identification and quantification of transcript isoforms[76,96]. Furthermore, it has been hypothesized that reverse transcription of full-length RNA molecules is impeded by cell or nucleus fixation which is required for certain single-cell protocols, leading to the increased appearance of truncated transcripts[80]. Both of these problems further exacerbate the sparsity of single-cell LR-RNA-seq data. Moreover, there has been a lack of comprehensive investigation into the extent to which single-cell data processing methods can effectively generalize to single-cell LR-RNA-seq that address the specific research questions that can be answered by the technology.

1.3 Conclusions

Long-read sequencing platforms have enabled the full-length characterization and quantification of transcriptomes, which is important due to the profound impact that alternative transcript usage can have in development and disease. Current studies that use LR-RNA-seq have typically been restricted to smaller scale projects that seek to understand the difference between transcript usage in a limited number of biological contexts or samples. My project aims to characterize the diversity of transcription from full-length LR-RNA-seq data across samples, genotypes, and cell types.

In Chapter 2, I discuss Swan, which is a software tool that I developed for downstream analysis and visualization of full-length transcriptomes. The Swan manuscript was published in *Bioinformatics* in 2020 and showcases its capabilities on transcriptomes derived from PacBio LR-RNA-seq data from the HepG2 and HFFc6 cell lines from ENCODE. I used Swan to highlight the statistical testing features as well as the unique visualizations that Swan produces. I have further updated the chapter to demonstrate newer features I added to Swan by applying it to a more complex dataset comparing wild type and 5xFAD mice from different sexes, ages, and brain tissues.

In Chapter 3, I present the work we did to characterize transcriptional diversity across the ENCODE4 PacBio LR-RNA-seq datasets from human and mouse. We developed a computational pipeline that allows us to call unique transcripts with the same exon junction chain that use different transcript start sites (TSSs) or different transcript end sites (TESs) and name transcripts according to their structural content. With this approach, we observed that protein-coding genes have a median of two unique predominant transcript isoforms, which represent the most highly expressed transcript per gene in each sample. We introduced a novel method of describing transcriptional diversity that captures the contributions of alternative TSS and TES usage as well as alternative splicing and showed that this method can be used to compare transcriptional diversity across a range of biological contexts. We further demonstrate that using this metric, for orthologous protein coding genes in human and mouse, transcriptional diversity is not well conserved between the species. This work was posted as a BioRχiv preprint in May 2023 and is currently in revision.

In Chapter 4, I discuss the application of our single-nucleus LR-RNA-seq assay, LR-Split-seq, to the prefrontal cortex of two genetically distinct mouse strains, C57BL6/J and CAST/EiJ.

We were able to find instances of transcript isoform switching between the different strains which were particularly pronounced in both the GABAergic and glutamatergic neuronal subpopulations. This work was done as a part of our involvement in the Impact of Genetic Variation on Function (IGVF) consortium and will be used to write a manuscript with our IGVF collaborators before the end of 2023.

Finally, in Chapter 5, I propose future directions for this work as well for the field of full-length transcriptomics. I believe this field is still truly in its nascent stages as I have watched the technology improve immensely over the last five years, and we have only begun to scratch the surface regarding the biological insights that we can glean from the massive amounts of data now being produced. Exploring the effect of genetic variation on alternative transcript usage, probing RNA modifications, and improving single-cell and nucleus LR-RNA-seq approaches are just a few of the topics that have yet to be thoroughly explored.

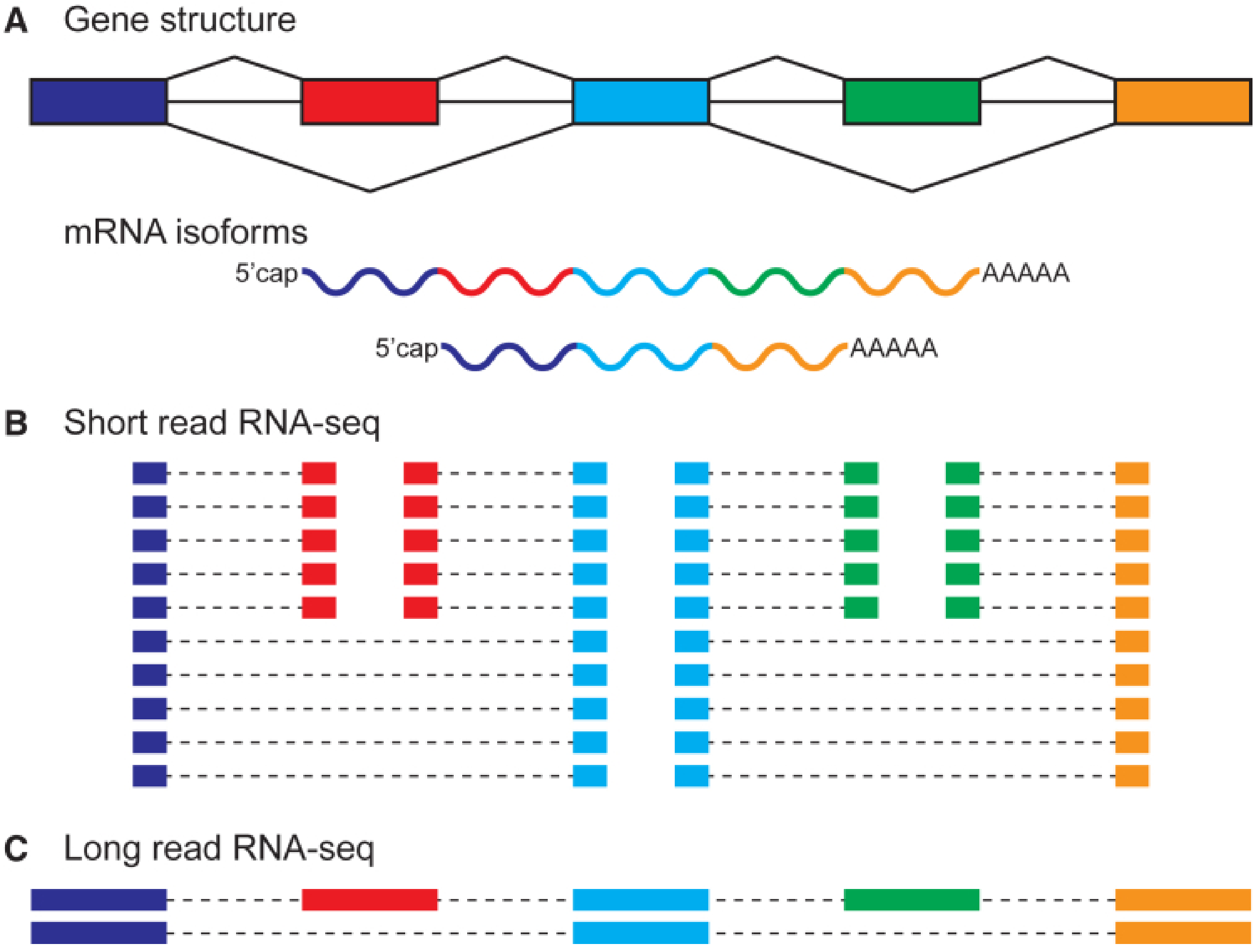


Figure 1.1: **Long-read RNA-seq captures full-length transcript isoforms. a,** Genomic DNA structure of an alternatively spliced gene and its resultant mRNA transcript isoform products. **b,** Representation of exon-exon junction containing short reads that would result from sequencing this gene. **c,** Representation of full-length long reads that would result from sequencing this gene. (Figure from Park et al.[5]).

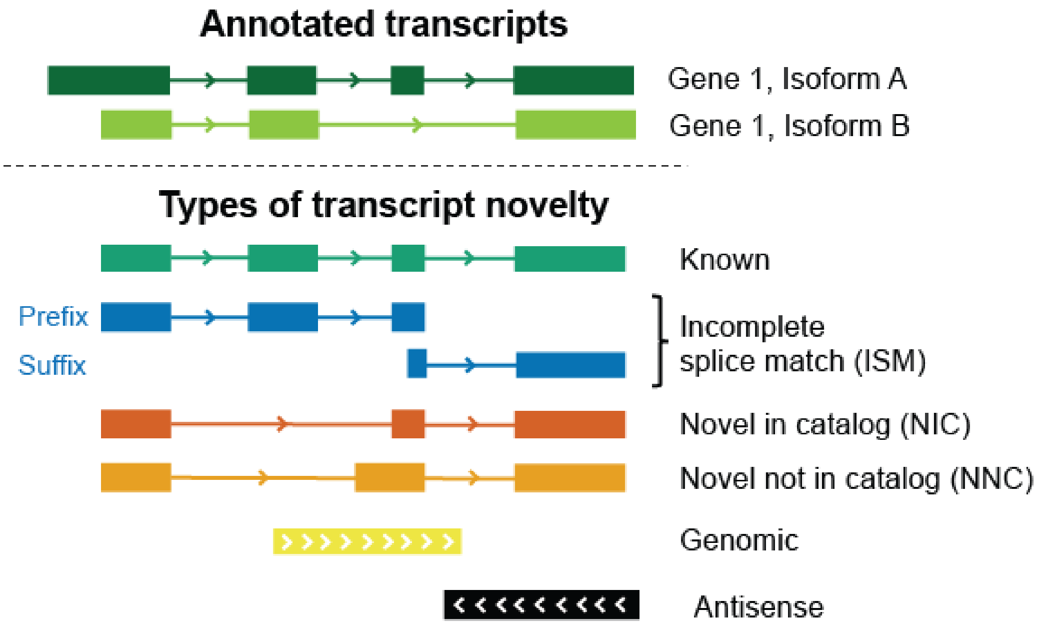


Figure 1.2: **Transcript novelty categories determined by comparison to the reference annotation.** (Figure from Wyman et al.[36]).

Chapter 2

Swan: a library for the analysis and visualization of long-read transcriptomes

Note: This chapter was published in *Bioinformatics* in 2020. It has been revised with additional data and text to reflect updates made to the software since then.

2.1 Abstract

Long-read RNA-sequencing technologies such as PacBio and Oxford Nanopore have discovered an explosion of new transcript isoforms that are difficult to visually analyze using currently available tools. We introduce the Swan Python library, which is designed to analyze and visualize transcript models. Swan finds 4,909 differentially expressed transcripts between cell lines HepG2 and HFFc6, including 279 that are differentially expressed even though the parent gene is not. Additionally, Swan discovers 285 reproducible exon skipping

and 47 intron retention events not recorded in the GENCODE v29 annotation. We additionally apply Swan to a more complex dataset from control and 5xFAD mice. We find evidence of TSS, isoform, and TES switching between the genotypes. The Swan library for Python 3 is available on PyPi and on GitHub.

2.2 Introduction

Alternative splicing plays a critical role in many biological processes and disease states. However, standard genomic assays have difficulties capturing the comprehensive spectrum of full-length alternative splicing due to the limitations of short reads in reconstructing transcript isoforms[98]. Long-read sequencing platforms such as PacBio and Oxford Nanopore have led to an explosion in discovery of transcript isoforms that were impossible to assemble with short reads. Current transcript model visualization tools such as Sashimi or LeafCutter plots are primarily designed for visualizing short reads rather than leveraging full-length isoforms[99,100]. Furthermore, these tools display transcript isoforms on a genomic scale which complicates interpreting and distinguishing similar transcripts from one another.

We introduce the Swan Python library, which is designed to analyze and visualize transcript models. Swan's graphical model approach allows the user to visually distinguish between transcript models and to identify novel exon skipping and intron retention events commonly missed by short reads. Furthermore, Swan incorporates statistical models to detect differential gene and transcript expression, enabling quantitative comparison of full-length transcript models in different biological settings. While Swan was designed with long reads in mind, it can support any transcriptome from a properly formatted GTF. We demonstrate Swan's utility by applying it to full-length PacBio transcriptome data from the HepG2 and HFFc6 human cell lines, which are publicly available on the ENCODE portal at the accessions listed in Table 2.1.

We furthermore demonstrate the updated utility of Swan by applying it to a more complex dataset consisting of 5xFAD and wildtype (C57BL6/J) mouse samples produced as part of MODEL-AD as processed in the ENCODE4 long-read RNA-seq dataset[101]. We find evidence of TSS, isoform, and TES switching between the genotypes.

2.3 Materials and methods

2.3.1 Input and data structure

Transcriptome input and representation

Swan works by processing transcript models from either GTF files or from a TALON database[36] into a SwanGraph data structure consisting of a series of data frames and a graph. Each unique genomic coordinate, exon, intron and transcript is recorded in the data frames and used to construct a graphical representation of the transcriptome. In Swan, this combination of data structures is known as a SwanGraph. Genomic locations are represented as nodes. Introns and exons are represented as edges between nodes, and a full transcript is represented as a path that traverses the nodes and edges present in the transcript. This flexible scheme makes it possible to add additional datasets and track which nodes, edges and full transcripts are present in each.

Transcript expression and metadata input and representation

Expression data can also be added for each dataset from a transcript ID-indexed counts matrix TSV file. Swan will automatically calculate the expression and relative usage (percent isoform[57]) of each transcript and its constituent features (transcript start sites, transcript end sites, and exon junction chains), and store these values in AnnData[102] format, which is a commonly-used format to represent single-cell data in sparse matrix format as well as metadata, which can be provided in TSV format. Using the AnnData structure facilitates fast calculations, low memory and storage footprint, storage of complex hierarchical metadata, and direct compatibility with Scanpy plotting and analysis features[102]. In the case that a transcriptome from Cerberus[103] is added, Swan will automatically use the TSSs, exon junction chains, and TESs assigned by Cerberus to calculate expression and usage of these features. Optionally, the user can also add gene-level expression information.

Compatibility with single-cell data

Swan's sparse data representation and usage of the AnnData data structure allows for compatibility with single-cell and nucleus datasets as well. For ease of use, users can add abundance and metadata information directly from a preexisting AnnData object, which bypasses the dense-matrix representation which in many cases is storage and memory prohibitive. By default in single cell mode, Swan does not calculate percent isoform usage of each isoform in a given cell; instead only opting to compute these statistics across a particular metadata condition for downstream analytical tasks.

2.3.2 Visualization

SwanGraph visualizations

Swan's plots directly correspond to the graphical representation present in the SwanGraph; the nodes correspond to genomic coordinates, and the edges represent exons (regions between genomic coordinates present in the transcript) and introns (regions between genomic coordinates not present in the transcript) (Fig. 2.1). Nodes and edges are colored according to their role in the gene or transcript. Nodes that are used as transcription start sites (TSSs) are colored blue, nodes that are used as transcription end sites (TESs) are colored orange, and internal nodes that are traversed between the start and end of a transcript are colored yellow. Nodes with more than one classification (for example a node that can be used either as a TSS or an internal node) are preferentially colored by their most "unique" role in the gene (TES > TSS > internal). Exonic edges are colored green and intronic edges are colored pink. Nodes are spaced out evenly regardless of genomic location. This makes it easier to differentiate splicing events such as alternative 5'/3' splice sites that are only a few base pairs away from the canonical splice sites which are difficult to distinguish on a genomic scale. Instead of the difference between these splice sites consisting of a difference of just a few pixels such as in a genome browser-style representation, Swan's visualization uses an entirely different node to draw the user's attention to the heterogeneity of splicing (Fig. 2.1).

Swan offers different visualization options. Gene summary graphs provide a high-level overview of the splicing complexity of a specific gene where every node and edge present in a gene is brightly colored (Fig. 2.2a). Transcript path graphs show the nodes and edges traversed by a specific transcript isoform in bright colors through the rest of the gene graph, where unused nodes and edges will be grayed-out (Fig. 2.2b). In both gene summary and transcript path graphs, options exist to draw attention to nodes/edges that are not present in the annotation (novel nodes/edges) or to nodes/edges that are seen in a specific dataset (Fig. 2.2b).

Browser-style plots

Transcript paths can also be plotted using a traditional genome browser-style representation to provide a familiar reference. These browser-style models can also be plotted directly onto a figure axis alongside other figure elements. Swan can also plot BED regions that can be displayed at the same genomic scale as the transcript models to provide additional context about the genomic region that the transcripts are located in (Fig. 3.15d).

Swan reports

Swan can also generate a PDF report for a given gene which will plot the transcript path representation either in the nodes and edges representation or using the browser-style plots for each transcript isoform in the gene, along with the presence or absence of each isoform in the datasets that have been added to the SwanGraph. If abundance information has been provided, Swan can add a heatmap to illustrate expression levels or percent isoform usage of each transcript. There are many other additional options to tune the appearance of the report, including the ability to aggregate information across different metadata conditions, choose which transcript identifier is displayed, change the display order of the datasets, and display multiple categories of metadata information using provided color mappings.

2.3.3 Analysis

Differential gene and transcript expression testing

Swan can perform differential expression tests using diffxpy[102]. Additionally, differential expression tests can be run directly using Scanpy functions on the AnnData objects stored in the SwanGraph.

Isoform switching testing

Unlike differential expression tests, isoform switching tests look only at the relative usage of a transcript within a gene across conditions with no regard to absolute expression values. This test therefore can find instances where the relative prevalence of a transcript is different between conditions. Swan's isoform switching test is implemented according to the strategy described in Joglekar et al.[57], and can be performed on the whole transcript, TSS, TES, or exon junction chain level.

Novel intron retention and exon skipping event identification

Swan can detect novel exon skipping and intron retention events using the graphical transcriptome representation. For intron retention, Swan finds novel exonic edges that span intronic edges. For exon skipping, Swan finds novel intronic edges that span exonic edges (Fig. 2.1b).

2.4 Results

Pilot Swan analysis

We obtained mapped PacBio RNA-seq datasets from the ENCODE portal for two replicates of HepG2 and three replicates of HFFc6. We used TALON v5 to call transcript isoforms and filter novel ones for high reproducibility for 22,857 transcript models in HepG2 and 28,814 in HFFc6. We then used TALON to obtain GTFs and abundance files for each dataset. We fed the resulting transcript models and abundance information into Swan, along with the GENCODE v29 annotation[104].

Using diffxpy, Swan identified 4,009 differentially expressed genes and 4,909 differentially expressed transcripts across cell lines. Of the differentially expressed transcripts, 279 belong to a parent gene that is not differentially expressed and are therefore candidates for isoform switching analysis. In addition, Swan found 285 novel exon skipping events and 47 novel intron retention events. The gene *ADRM1* is a candidate isoform switching gene based on the aforementioned criteria. Moreover, the *ADRM1* transcript that was called as differentially expressed contains a novel intron retention event (Fig. 2.2, 2.3).

Swan analysis of the 5xFAD Alzheimer's mouse model

To further demonstrate the utility of Swan, we applied it to a more complex dataset to compare the transcriptomes between wild-type samples and from a mouse model of Alzheimer's disease (5xFAD). From the ENCODE4 long-read RNA-seq dataset[103], we obtained processed data for all of the WT (C57BL6/J) and 5xFAD brain samples in mouse (Table 2.1). Between genotypes, we found 696 TSS switching events, 1,676 isoform switching events, and 965 TES switching events, as well as 1,466 novel exon-skipping events and 8,402 novel intron retention events. One isoform switching gene was *Csf2ra*, which is a macrophage colony stimulating factor subunit that has previously been known to engage in alternative splicing resulting in functionally distinct protein isoforms (Fig. 2.4)[105]. This is a particularly interesting finding as *Csf*-family genes, namely *CSF1R*, have been implicated in maintenance of microglial homeostasis in the brain[106]. Mutations in and dysfunction of *CSF1R* have been implicated in neurological and neurodegenerative diseases such as Alzheimer's disease[106,107] and overexpression of *Csf2* in mice has been associated with microgliosis[108]. We additionally called differentially expressed transcripts and genes using PyDESeq2[109] directly with our SwanGraph AnnData expression representations, and found 520 and 545 differentially expressed transcripts and genes respectively between genotypes.

2.5 Discussion

As long-read RNA-seq as a sequencing assay has recently matured immensely, there arises a clear need for computational tools to facilitate analyses that harness the full length nature of the data. Swan provides a such platform for deeply exploring full-length transcriptome data. Its intuitive visualizations and flexible analysis tools enable discovery of novel splicing events and differential transcript usage.

Using Swan, we showed that we were able to find differentially expressed genes and transcript; transcript, TSS, and TES usage changes; and novel intron retention and exon skipping events in two distinct datasets. Notably, we found an transcript isoform switching event in *Csf2ra* between 5xFAD and WT mice, demonstrating that a key factor in maintaining microglial homeostasis changes predominant transcript isoforms between conditions.

In particular, the SwanGraph data representation provides a clear avenue for exploring transcriptomes at different granularities. By representing each transcript as a series of splice sites or splice junctions, the relative abundances of splice sites or junctions can too be analyzed in isolation, as well as TSSs and TESs. Furthermore, the data representation would be ideal for studying the coordination of different alternative splicing events in conjunction, which are known to undergo coordinated inclusion or exclusion even when far apart on the genomic scale[76]. Additionally, the data representation naturally lends itself to graphical-based predictive tasks. In this case, edge weights derived from the usage of each exon or intron in the SwanGraph could be used to predict potential novel isoforms that might arise from the exons and introns already annotated.

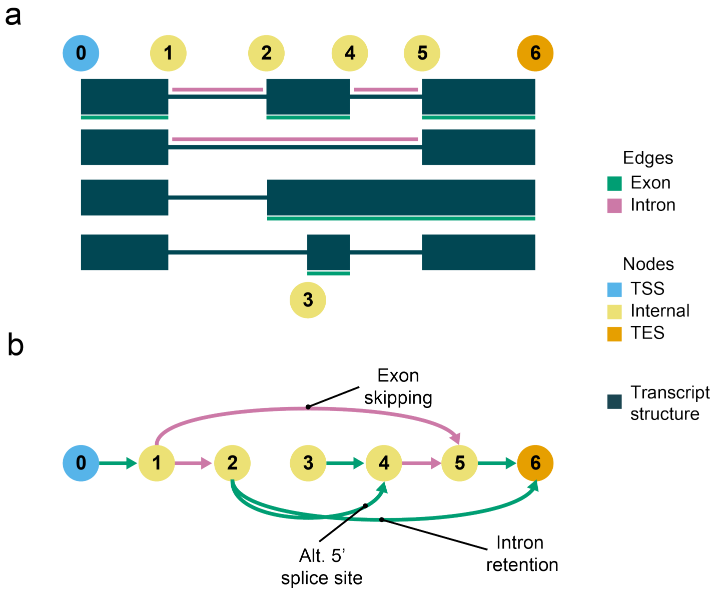


Figure 2.1: **Overview of Swan's graphical transcriptome representation a,** Browser-style models of 4 transcripts from the same gene. Unique nodes and edges as reflected in the SwanGraph are labeled. **b,** SwanGraph representation of transcripts in **a**, with specific edges labeled by the type of alternative splicing event they illustrate.

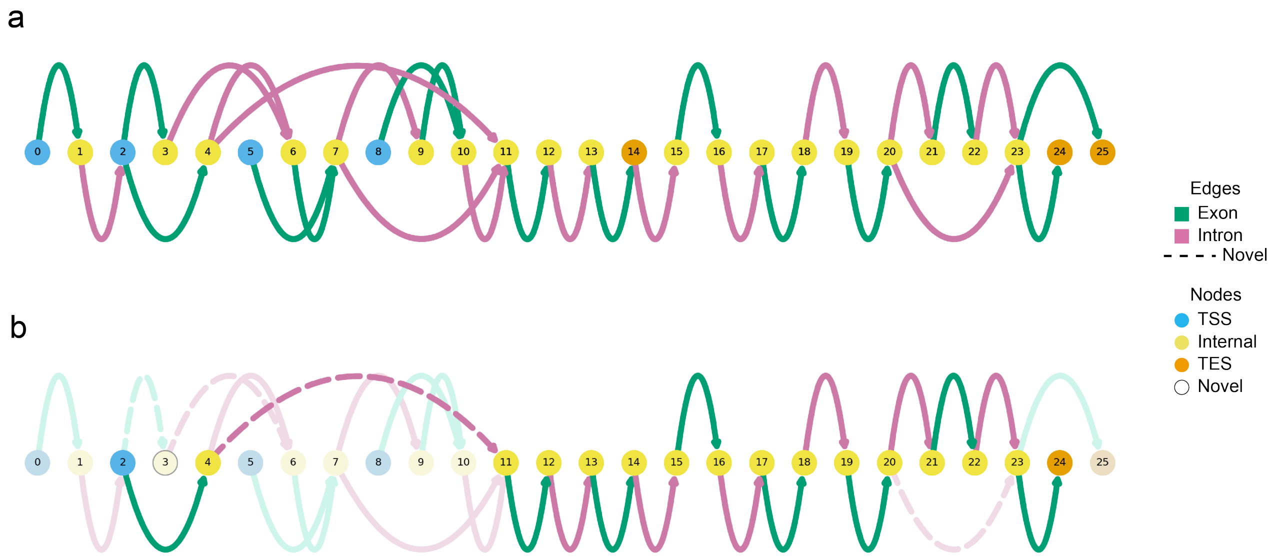


Figure 2.2: **SwanGraph plots for transcripts from *ADRM1*. a,** Gene summary plot representing all known (GENCODE v29) and novel transcripts called from *ADRM1* in HFFc6 and HepG2 data. **b,** SwanGraph transcript path plot for novel *ADRM1* transcript ENCODET000312957, which contains an intron retention event.

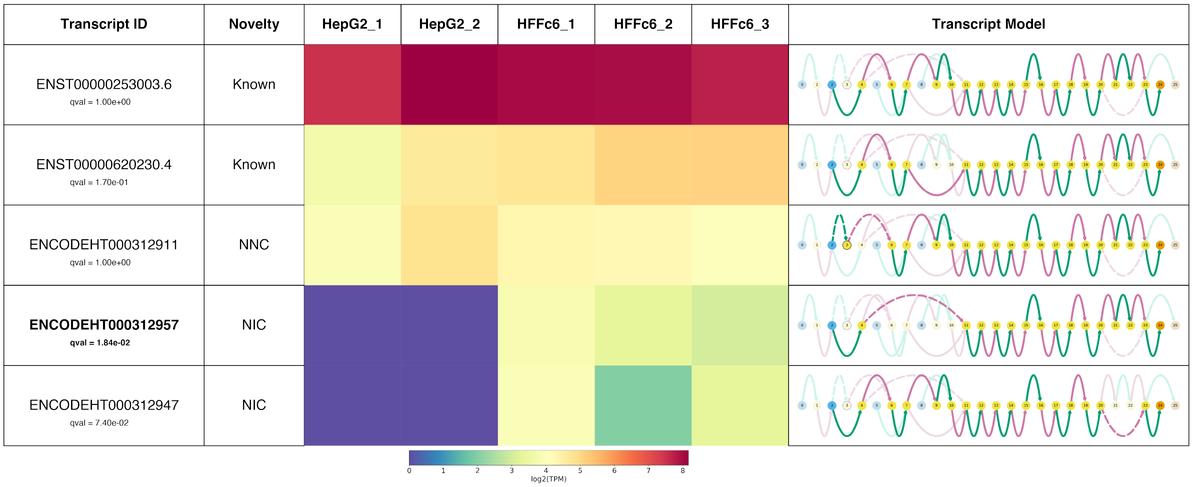


Figure 2.3: **Swan report for transcripts from *ADRM1*.** Heatmap shows expression level for each transcript. Bolded and asterisked models are differentially expressed transcripts between human cell lines HepG2 and HFFc6 ($q \leq 0.05$). Transcript novelty categories are determined by TALON. Novel splice sites are outlined nodes and novel combinations of splice sites are dashed edges. TSS nodes are blue and TES nodes are orange. Exonic edges are green and intronic edges are pink.

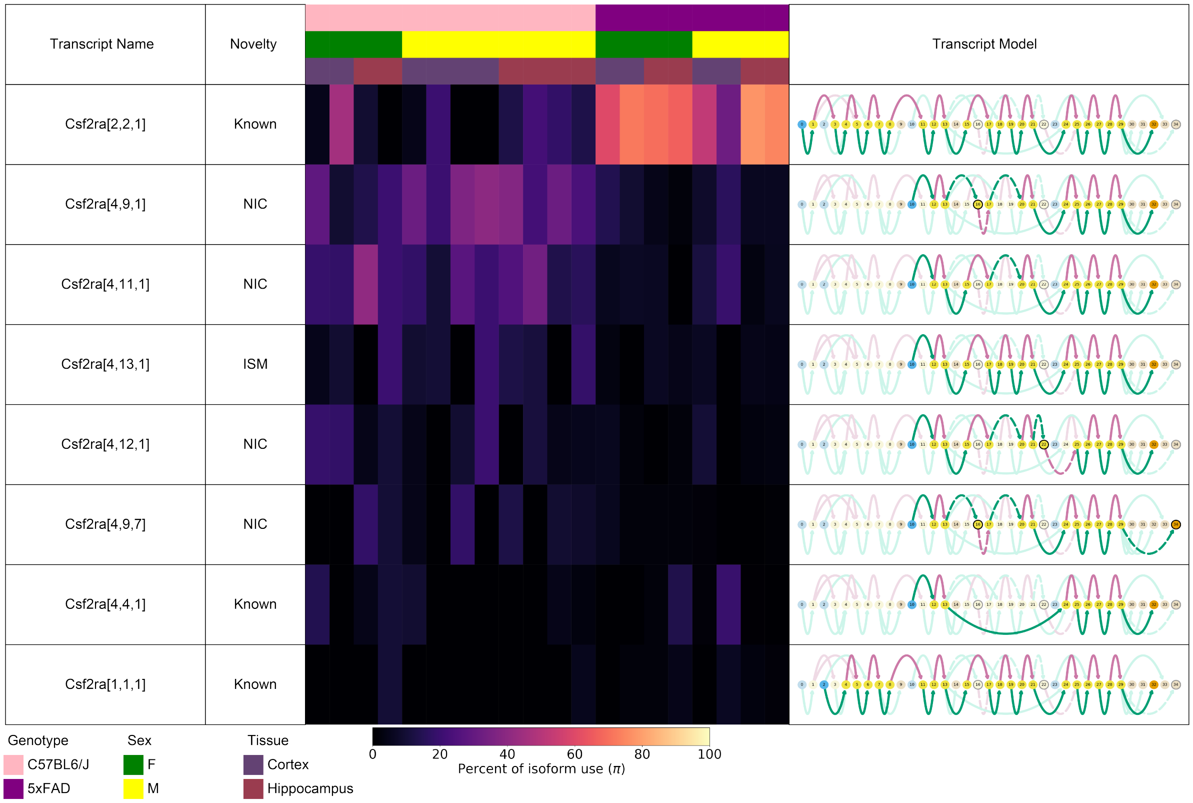


Figure 2.4: **Swan report for transcripts from isoform switching gene *Csf2ra*.** Heatmap shows percent isoform values for each transcript. Novel splice sites are outlined nodes and novel combinations of splice sites are dashed edges. TSS nodes are blue and TES nodes are orange. Exonic edges are green and intronic edges are pink.

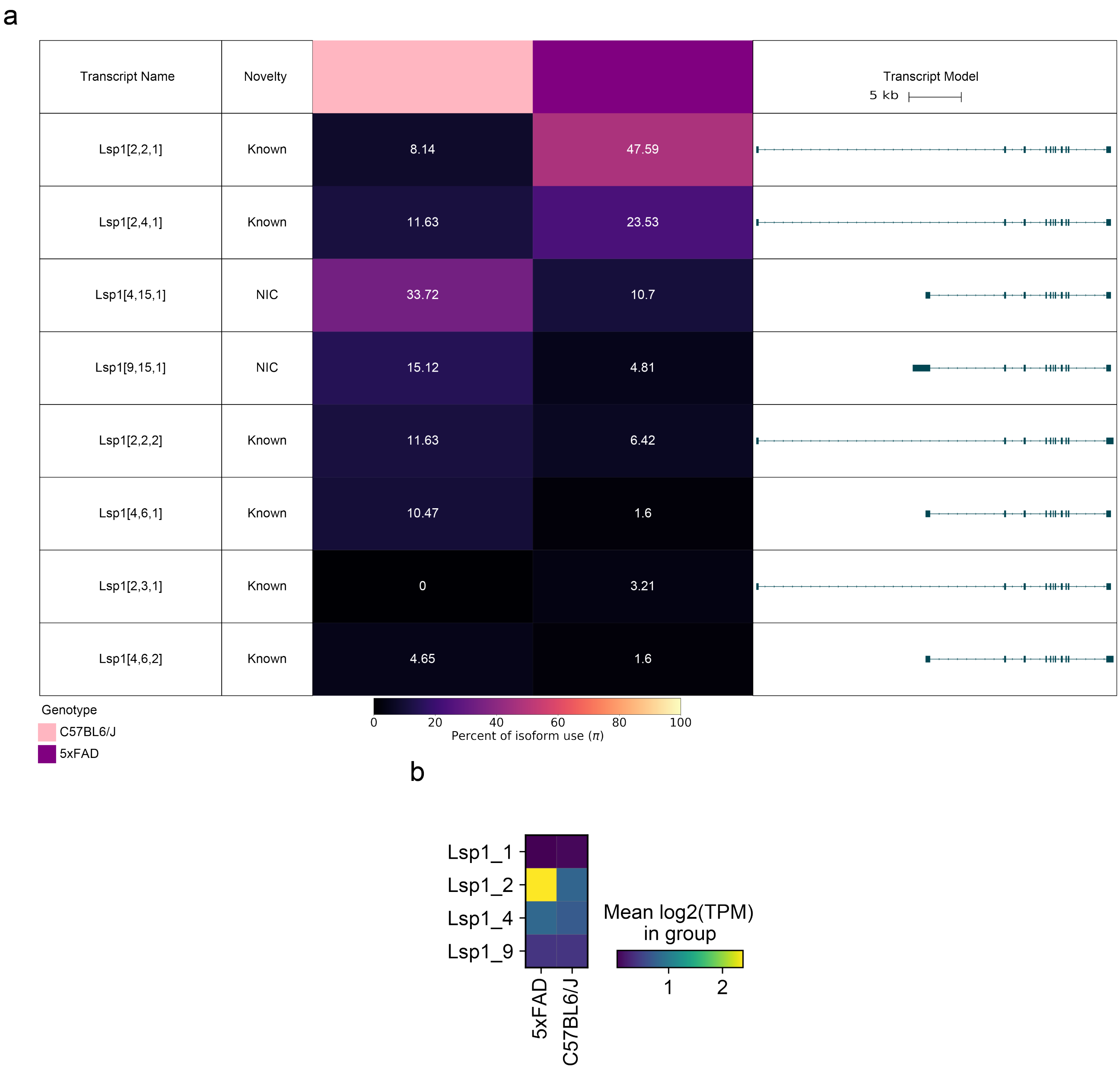


Figure 2.5: **TSS switching gene *Lsp1*. a,** Swan report showing the transcript structure and overall percent isoform values for the C57B6/J and 5xFAD genotypes for each *Lsp1* transcript. **b,** Scanpy matrixplot showing the expression of each TSS separately by genotype.

ENCODE accession	Sample description	Swan analysis
ENCSR834DQL	HepG2	Pilot
ENCSR902GAF	HFFc6	Pilot
ENCSR131CES	Cortex WT Female	MODEL-AD
ENCSR644GDT	Cortex WT Male	MODEL-AD
ENCSR340GWV	Cortex WT Male	MODEL-AD
ENCSR411ZHA	Hippocampus WT Female	MODEL-AD
ENCSR926OGQ	Hippocampus WT Male	MODEL-AD
ENCSR214HSG	Hippocampus WT Male	MODEL-AD
ENCSR280VKU	Cortex 5xFAD Female	MODEL-AD
ENCSR674BKT	Cortex 5xFAD Male	MODEL-AD
ENCSR060OTU	Hippocampus 5xFAD Female	MODEL-AD
ENCSR404AEI	Hippocampus 5xFAD Male	MODEL-AD

Table 2.1: ENCODE experiment accessions for data used in this study.

Chapter 3

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

3.1 Abstract

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 multitranscript 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 diversification 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.

3.2 Introduction

Most mammalian genes produce multiple distinct transcript isoforms[5]. 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[6–9]. 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[21,110], 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) project[50]. 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[23,111]. 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[24,26]. 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 GENCODE[21], 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[73,112,113]. Short-read RNA-seq has also been used by many groups to comprehensively characterize TSS usage[114], splicing[64], and TES usage[115], 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.

3.3 Results

The ENCODE4 RNA dataset

This LR-RNA-seq study profiled 81 tissues or cell lines by using the PacBio sequencing 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 timepoints in mouse, they represent 49 unique tissues or cell types across human and mouse (Fig. 3.1a, Fig. 3.2). In addition, we sequenced matching human short-read RNA-seq (Fig. 3.2c) and microRNA-seq (Fig. 3.3, 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. 3.1b, Fig. 3.4a). 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.

3.1c, Fig. 3.4b).

For each LR-RNA-seq dataset, we first mapped the reads using Minimap2[33] and corrected non-canonical splice junctions and small indels using TranscriptClean[38], after which we ran TALON[36] and LAPA[41] 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. 3.1d) 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. 3.1e). 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. 3.4c). 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. 3.1f).

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 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. 3.1g, 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. 3.4d). 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. 3.4d-g). In mouse, we observe similar read length profiles, sample separation, and gene detection patterns (Fig. 3.4h-j), including detection of 84.9% of annotated GENCODE vM25 protein coding genes at ≥ 1 TPM (Fig. 3.4h). 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[75,115] (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. 3.5a). 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. 3.5b, Fig. 3.6a). 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) of transcripts 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 $\geq$ 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 $\geq$ 1 TPM in at least one library from polyA genes) in our dataset (Fig. 3.6b-d, Fig. 3.7a-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. 3.6b-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. 3.6b-d, Fig. 3.7a-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. 3.5a, Fig. 3.7d-e, Supplementary methods).

We further annotated the novelty of ECs compared to GENCODE using the nomenclature from SQANTI[42]. 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. 3.7f). 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. 3.8, 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. 3.6e-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. 3.6h). 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. 3.9a, Fig. 3.1f). Of the remaining protein coding transcripts with novel ECs, 53.0% are predicted to have complete 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. 3.6i). 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. 3.6j, 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. 3.9b, 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 multiexonic 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. 3.9c-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. 3.9e-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. 3.12, Supplementary results).

Predominant transcript structure differs across tissues and cell types

Different multiexonic 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. 3.6k) whereas we detect 18 transcript isoforms that vary on the basis of their TSSs, ECs, and TESs for *PKM* (Fig. 3.6l).

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 the percentage of reads from a gene that come from that transcript (pi - percent isoform)[57] 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. 3.6m). 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. 3.6n). 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. 3.10a, Fig. 3.13). 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. 3.10a, Fig. 3.13). 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. 3.10a). Representing these numbers as proportions allows us to plot them as coordinates in a two-dimensional gene structure simplex (Fig. 3.10b, Fig. 3.13). 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. 3.10a-b, Fig. 3.13, 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. 3.10c-j, Supplementary methods). The observed and observed major gene triplets describe the diversity of transcription in each gene across all samples in the 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. 3.10f, 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. 3.10g). 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. 3.10f-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.

We performed functional enrichment testing for genes in each sector from our observed major set to determine shared characteristics of genes in each sector (Supplementary methods) (Fig. 3.11). Most notably, we found that genes in the TSS, TES, and mixed sectors were enriched for DNA binding activity, demonstrating that DNA binding genes exhibit a large range of transcript diversity behavior. The splicing, TES, and mixed sectors were also enriched for RNA binding. The apparent diversity of transcripts from RNA binding genes is unsurprising given the predisposition of RNA binding proteins such as splicing factors to themselves be enriched for alternative splicing[17]. In fact, when limiting analyzed genes to just splicing factors, we see that only 7.0% of genes fall in the simple sector as compared to the 33.9% when considering all genes, demonstrating a clear tendency of splicing factors toward higher transcript diversity (Supplementary methods). Finally, the simple sector yielded several functions related to different signaling methods, such as hormone and cytokine activity. This demonstrates that the flexibility of transcription is limited for genes involved in signaling.

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. 3.10k).

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[116,117]. 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. 3.10k). 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. 3.10l-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. 3.10d, Fig. 3.6j, Fig. 3.14a-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. 3.6e-g, j), on the sample level, most genes have far fewer triplet features and transcripts; with a particularly pronounced difference for the TSS (Fig. 3.14a-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. 3.14e-l, 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 protein[118], 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. 3.14m). This drives the sample-level centroid behavior into the mixed sector (Fig. 3.14n). In contrast, in heart and muscle *ACTA1* expresses 18 and 15 transcripts respectively, which all differ on the basis of their ECs (Fig. 3.14m-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 matrix[119], 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. 3.15a). 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. 3.15b). By contrast, the four transcripts from the transcription factor *CTCF* expressed in lung use three TSSs but only one TES (Fig. 3.15c-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*[120] 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. 3.15e-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. 3.15g). 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 others[50]. 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. 3.15h). 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. 3.15i-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. 3.16a-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. 3.16e-h). Overall, for most gene / library combinations, the MANE transcript or triplet feature is the predominant one (Fig. 3.16i-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. 3.10f-j, Fig. 3.17a-e, Fig. 3.18a-c). For the mouse observed and observed major gene triplets, we see relatively similar percentages of genes in each sector (Fig. 3.17a-e, Fig. 3.18a-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. 3.18d). 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* family[121]. Human *ARF4* sample-level gene triplets are nearly always splicing-high whereas mouse *Arf4* sample-level gene triplets are mainly TES-high (Fig. 3.17f-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. 3.17h-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. 3.17j). 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. 3.17k). Thus, we find substantial differences in splicing diversity for orthologous genes between human and mouse.

3.4 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[71,122]. We also found that for a substantial number of 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 capture[123]. 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[35,42,44,96,124]. 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. *COL1A1*, 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 elements[125]. 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.

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

3.5 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. 3.3a). Overall, microRNAs are more sample-specific than known genes in both short and long-read RNA-seq (Fig. 3.3a, Fig. 3.1c, f, Fig. 3.4b). 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. 3.3b). 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. 3.3c), yet the most microRNAs expressed at ≥ 2 CPM (Fig. 3.3d). 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. 3.3b). 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. 3.3e). 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[126,127]. 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. 3.8a-b, Fig. 3.6b, Fig. 3.7a). We used a simple logistic regression model that incorporates expression, DNase-Hypersensitivity (DHS)[128], and length of our LR-RNA-seq observed TSSs (Supplementary methods, Fig. 3.8c-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. 3.8e), 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. 3.8f). 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 consortium[75]. We ran SUPPA2[129] 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. 3.12a), 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. 3.12b). 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. 3.12c). 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.

3.6 Supplementary methods

B6/Cast mouse tissue collection

Mouse tissues were harvested from C57BL/6J (RRID:IMSR_JAX:000664) x CAST/EiJ (RRID:IMSR_JAX:0
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.

RNA extraction, cDNA preparation, and PacBio sequencing

All details for preparation and sequencing of PacBio cDNA are detailed on the ENCODE portal. Relevant protocol document URLs for each library can be found in Tables S1-2 in the “document_url” column. RIN numbers, whether or not spike-ins were included, and the platform the library was sequenced on are also included.

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-b09c92215e`
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. 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. Briefly, reads are basecalled with CCS. We mapped reads to hg38 or mm10 using minimap2[33] with the following settings: `-ax splice -uf --secondary=no -C5`. We corrected for common long-read sequencing artifacts such as microindels and non-canonical splice junctions (sans those with support from the GENCODE v29 or vM21 reference annotation) with TranscriptClean[38]. The resultant BAM files are available as the "alignments" files for each file accession (specified in the "ENCODE_alignments_id" column of tables S1-2) 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 --ar 20` module and hg38 / mm10. A reference TALON database was initialized either from GENCODE v29 or vM21 using `talon_initialize_database --l 0 --5p 500 --3p 300`. Reads were annotated using TALON with the reference database as input and default settings. 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 --maxFracA 0.5 --minCount 5 --minDatasets 2` 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[41] on the BAM files output from TranscriptClean 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 ($\geq$ 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 $\geq$ 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

Cerberus fills several roles. Firstly, it provides a way to harmonize transcriptomes based on their TSS, EC, and TES content across diverse sources to determine transcript equivalence even if the annotations were not performed in the same way. This is particularly helpful with large datasets where it is time and memory prohibitive to annotate all long-read RNA-seq reads at the same time. Second, it allows for incorporation of data from external assays or references to either corroborate triplet features already in the Cerberus reference or to add new entries to the reference. This is particularly helpful for TSSs and TESs, where there are high-throughput assays that measure 5' or 3' end expression and identity, such as CAGE or PolyA-seq; and where there are already reference atlases that catalog these features such as the PolyA Atlas[115] and FANTOM[114]. Thirdly, in annotating transcripts according to their triplet features, it provides a way to distinguish transcripts from one another solely based on their names. Finally, Cerberus' downstream analysis tools leverage the triplet feature content of each transcript to classify genes based on their transcript structure diversity from a given set of transcripts.

Obtaining annotated TSS / TES regions from GTFs

Given a GTF, `cerberus gtf_to_bed` (Fig. 3.5a) 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 in the same gene are merged to ensure non-overlapping intervals. 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. 3.5a).

Assigning triplet features numbers

As part of both `cerberus gtf_to_ends` and `gtf_to_ics` (Fig. 3.5a), 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 APPRIS[49] 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 input BED file, `cerberus agg_ends` (Fig. 3.5a) 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. 3.5a) 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 separate aggregated TSS, EC, and TES files, `cerberus write_reference` (Fig. 3.5a) will write all three tables in a Cerberus reference h5 format. h5 is a well-supported (readers exist in both Python and R) and commonly used data structure that can store multiple tables. These results are stored in an additional table in the h5 file that maps each input transcript to the triplet features used.

Annotating a GTF with transcript triplets

Given an input GTF, `cerberus annotate_transcriptome` will determine which TSS, EC, and TES in the Cerberus reference set match each of the features in a transcript. For ECs, this relies on exact matching of the EC and the gene. For 5' and 3' ends, Cerberus looks for the closest TSS region from the same gene upstream or overlapping its TSS. For TESs, Cerberus looks for the closest TES region from the same gene downstream or overlapping its TES.

Updating 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. 3.5b). 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. 3.10a, Fig. 3.13). 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 previously annotated with `cerberus annotate_transcriptome`. `CerberusAnnotation.get_express` 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_tr` 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. 3.10a-b, Fig. 3.13).

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. 3.10a-b, Fig. 3.13).

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 such that it can still be accurately plotted in the simplex (Fig. 3.13).

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. 3.13). Jensen-Shannon distance is a metric on probability distributions[130]. For a given pair of gene structure simplex coordinates, the Jensen-Shannon distance is computed in Cerberus using Scipy[131] 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 data[114], and ENCODE PLS, pELS, and dELS cCREs. External datasets used to support TESs were obtained from ENCODE PAS-seq data, and the PolyA Atlas[115]. 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. 3.5a) 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_transciptome`, updated GTFs were generated with `cerberus replace_gtf_ids` with the update ends and collapse options used (Fig. 3.5b). 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. 3.5b).

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 (pi, 0-100) value that indicates what percentage of the gene's expression is derived from said transcript using Swan[56]. 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 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).

GO analysis of genes in each sector

We used GSEApy's Enrichr module[132,133] to determine the enrichment of gene function from the "GO_Molecular_Function_2021" database for genes in each of the five sectors as categorized by their observed major triplets. We selected GO terms shown for those that were in the top 5 results when sorted by most significant adjusted p-value.

Splicing factor sector analysis

We used this Biomart query to obtain a list of splicing factors. We determined each splicing factor gene's sector identity based on its observed major gene triplet, and compared the percentage of genes per sector overall in the observed major category versus the splicing factor observed major category.

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 using Scipy's `stats.zscore` function on all the distance values.

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. 3.5a) 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_transciptome`, updated GTFs were generated with `cerberus replace_gtf_ids` with the update ends and collapse options used (Fig. 3.5b). 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. 3.5b).

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 $\geq$ 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 Swan[56]. 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's[51] 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 BLASTP[134] 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 (v2.3[129]) 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 intersect[135] 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`

3.7 Supplementary tables

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

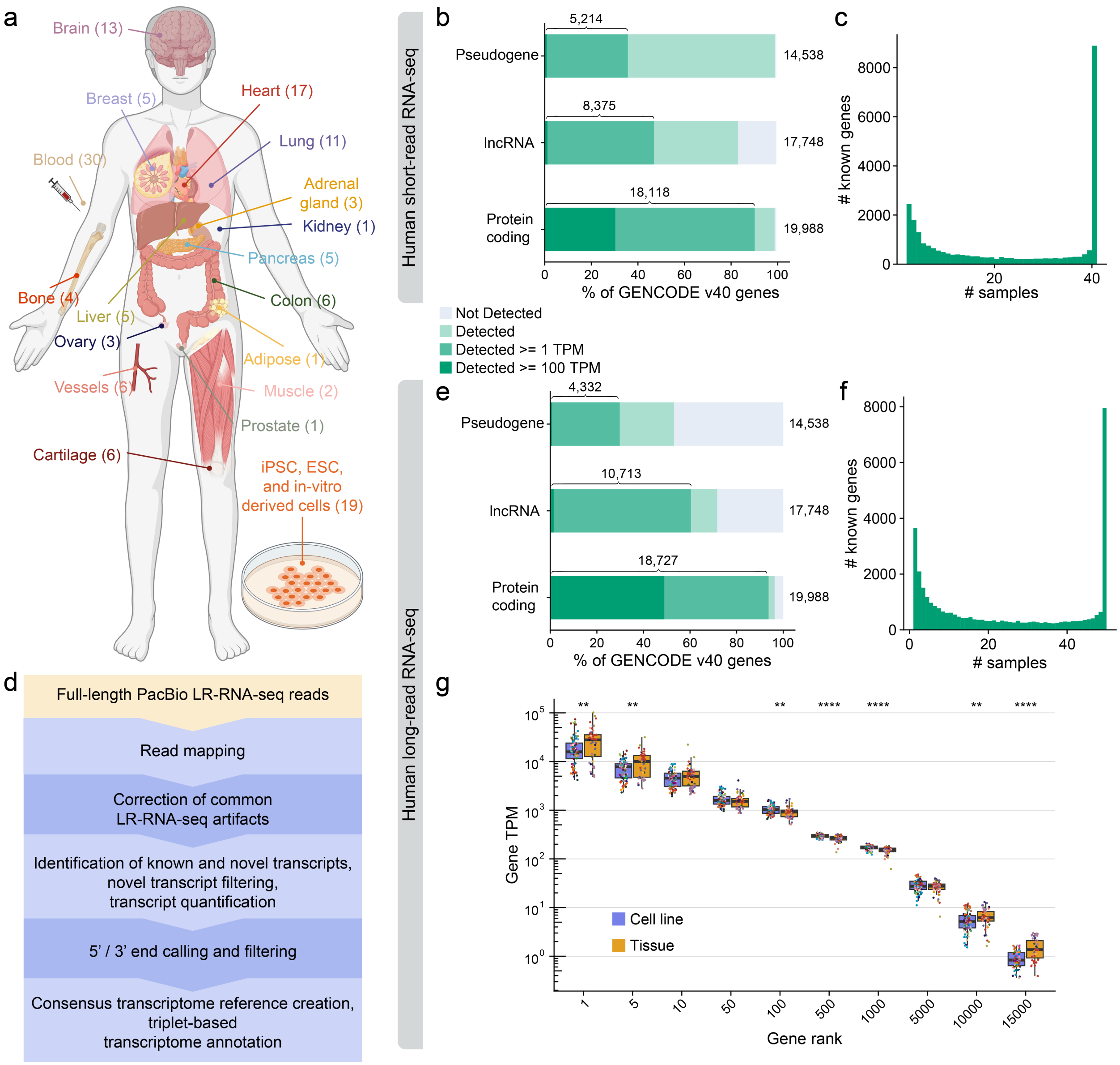


Figure 3.1: **Overview of the ENCODE4 RNA datasets. a,** Overview of the sampled tissues and number of libraries from each tissue in the ENCODE human LR-RNA-seq dataset. **b,** Percentage of GENCODE v40 polyA genes by gene biotype detected in at least one ENCODE short-read RNA-seq library from samples that match the LR-RNA-seq at $>$ 0 TPM, $\geq$ 1 TPM, and $\geq$ 100 TPM. **c,** Number of samples in which each GENCODE v40 gene is detected $\geq$ 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, $\geq$ 1 TPM, and $\geq$ 100 TPM. **f,** Number of samples in which each GENCODE v40 gene is detected $\geq$ 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 $\leq$ 0.05, **P $\leq$ 0.01, ***P $\leq$ 0.001, ****P $\leq$ 0.0001; Wilcoxon rank-sum test.

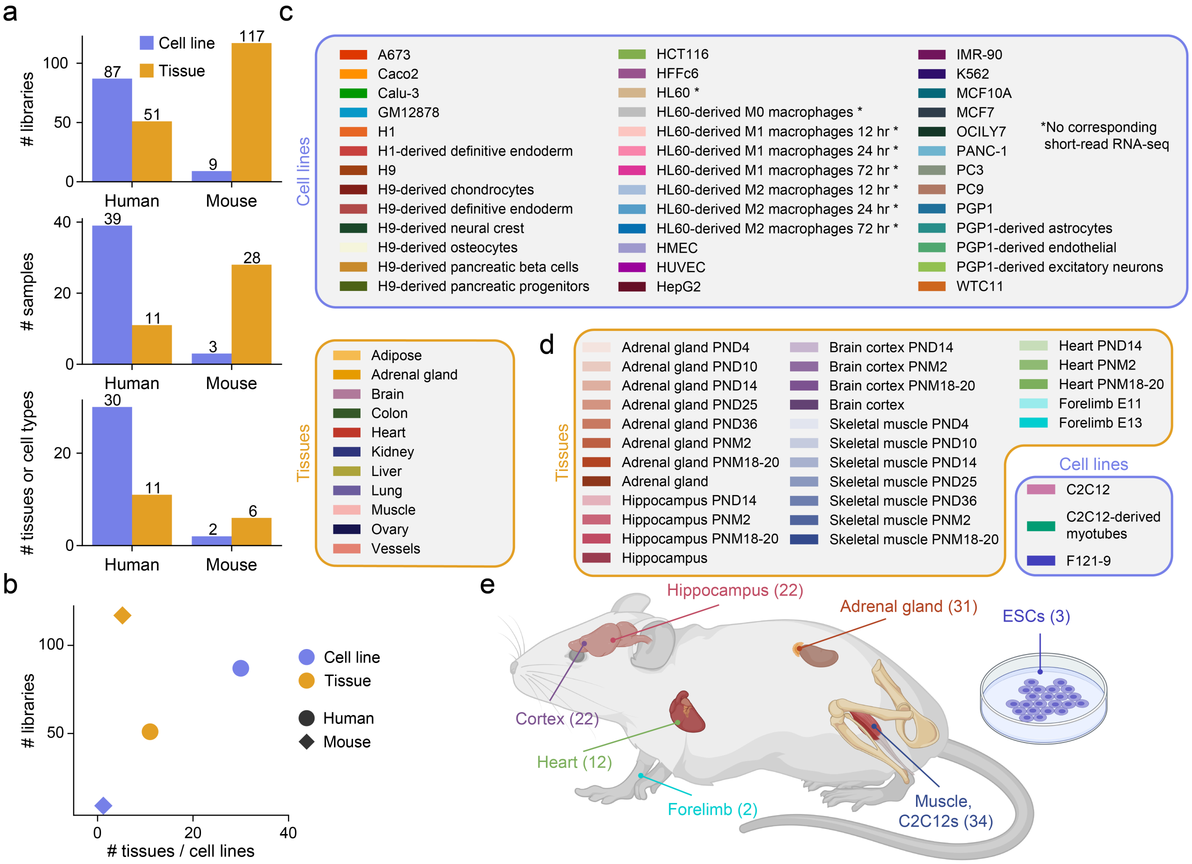


Figure 3.2: **Overview of the ENCODE4 LR-RNA-seq dataset. a,** From top to bottom, number of LR-RNA-seq libraries, samples (split by cell line / tissue identity as well as timepoint, when relevant), and unique tissues or cell types in the ENCODE LR-RNA-seq dataset split by species and tissue or cell 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.

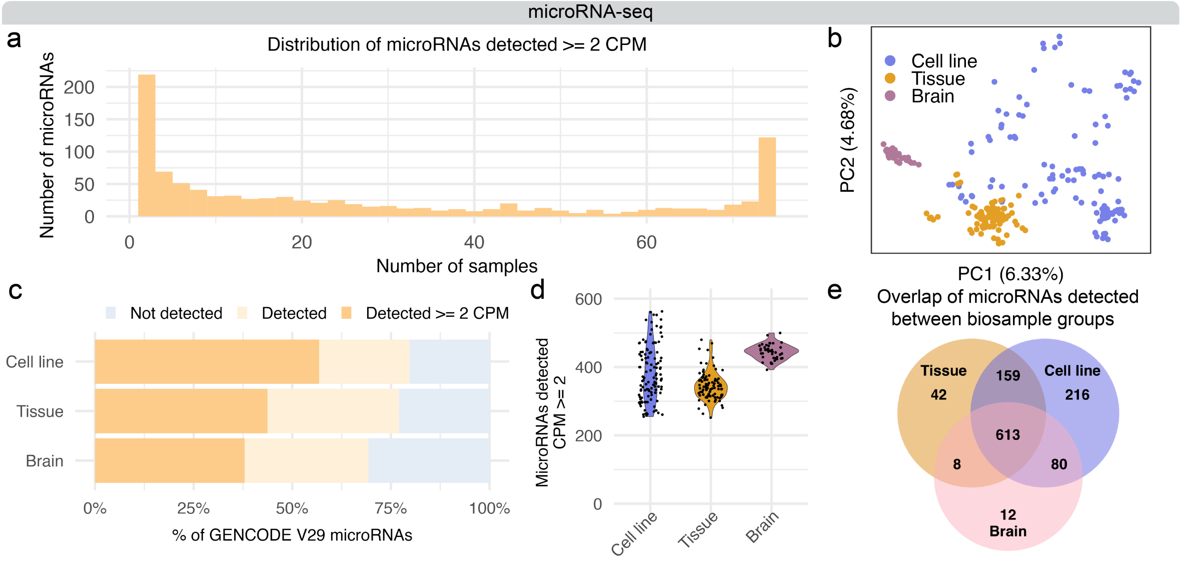


Figure 3.3: **Overview and detection of microRNAs in the ENCODE microRNA-seq dataset. a,** Distribution of GENCODE v29 mature microRNAs detected at CPM > 2 between cell lines, tissues, and brain tissue samples. **b,** PCA computed on microRNAs detected > 2 CPM in each human microRNA-seq library, colored 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.

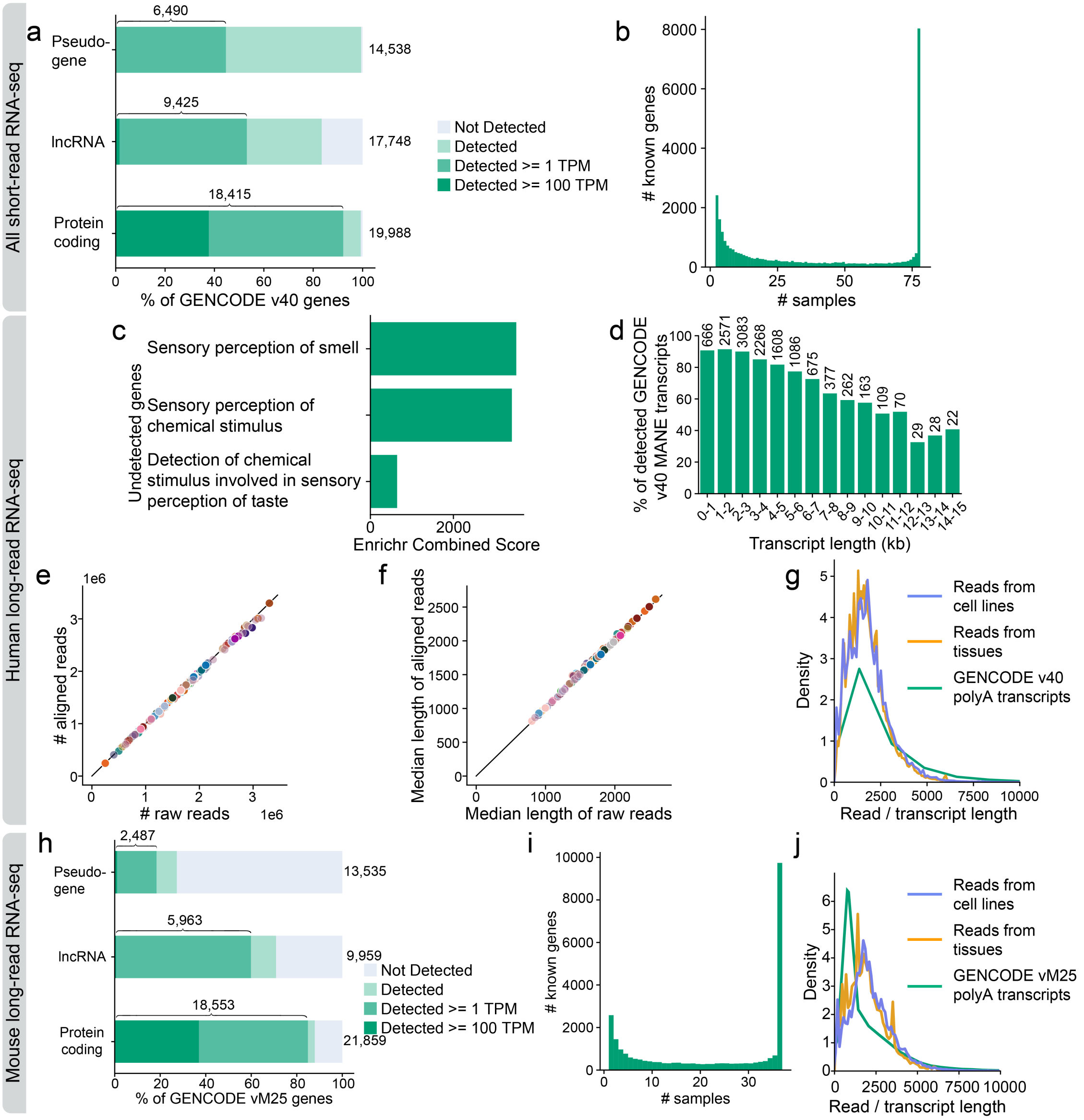


Figure 3.4: **Gene detection from short-read RNA-seq; gene detection, read length and alignment QC in both human and mouse LR-RNA-seq. a,** % of GENCODE v40 polyA genes by gene biotype detected in at least one ENCODE short-read library at various TPM thresholds. **b,** Number of samples in which each GENCODE v40 gene is detected in the ENCODE short-read RNA-seq dataset. **c,** Top 3 biological processes from undetected GENCODE v40 protein coding genes. **d,** % and number of detected GENCODE v40 MANE transcripts binned by transcript length. Restricted to genes $\geq$ 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 read alignment length in each human LR-RNA-seq library. **g,** Post-TALON read length profiles from human LR-RNA-seq data by tissue or cell line, and polyA transcript length profile from GENCODE v40. **h,** % of GENCODE vM25 polyA genes by gene biotype detected in at least one ENCODE mouse LR-RNA-seq library at various TPM thresholds. **i,** Number of samples that each GENCODE vM25 gene is detected in the ENCODE mouse LR-RNA-seq dataset. **j,** Post-TALON read length profiles from mouse LR-RNA-seq data by tissue or cell line, and polyA transcript length profile from GENCODE vM25.

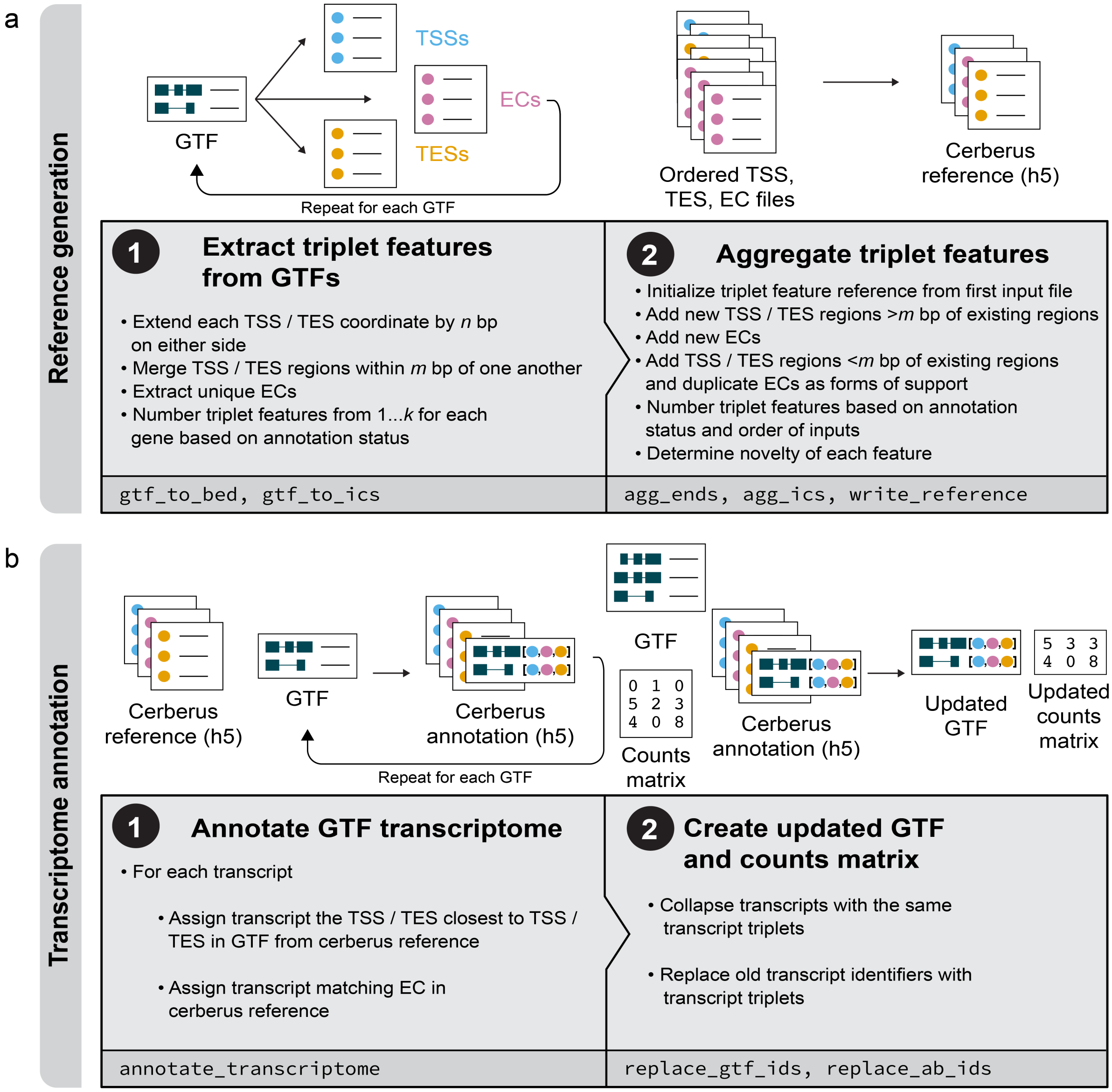

a
Reference generation
GTF
TSSs
ECs
TESs
Repeat for each GTF
Ordered TSS, TES, EC files
Cerberus reference (h5)
1 Extract triplet features from GTFs
• Extend each TSS / TES coordinate by n bp on either side
• Merge TSS / TES regions within m bp of one another
• Extract unique ECs
• Number triplet features from 1...k for each gene based on annotation status
gtf_to_bed, gtf_to_ics
2 Aggregate triplet features
• Initialize triplet feature reference from first input file
• Add new TSS / TES regions >m bp of existing regions
• Add new ECs
• Add TSS / TES regions <m bp of existing regions and duplicate ECs as forms of support
• Number triplet features based on annotation status and order of inputs
• Determine novelty of each feature
agg_ends, agg_ics, write_reference
b
Transcriptome annotation
Cerberus reference (h5)
GTF
Cerberus annotation (h5)
Repeat for each GTF
GTF
0 1 0
5 2 3
4 0 8
Counts matrix
Cerberus annotation (h5)
Updated GTF
5 3 3
4 0 8
Updated counts matrix
1 Annotate GTF transcriptome
• For each transcript
• Assign transcript the TSS / TES closest to TSS / TES in GTF from cerberus reference
• Assign transcript matching EC in cerberus reference
annotate_transcriptome
2 Create updated GTF and counts matrix
• Collapse transcripts with the same transcript triplets
• Replace old transcript identifiers with transcript triplets
replace_gtf_ids, replace_ab_ids

Figure 3.5: **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.

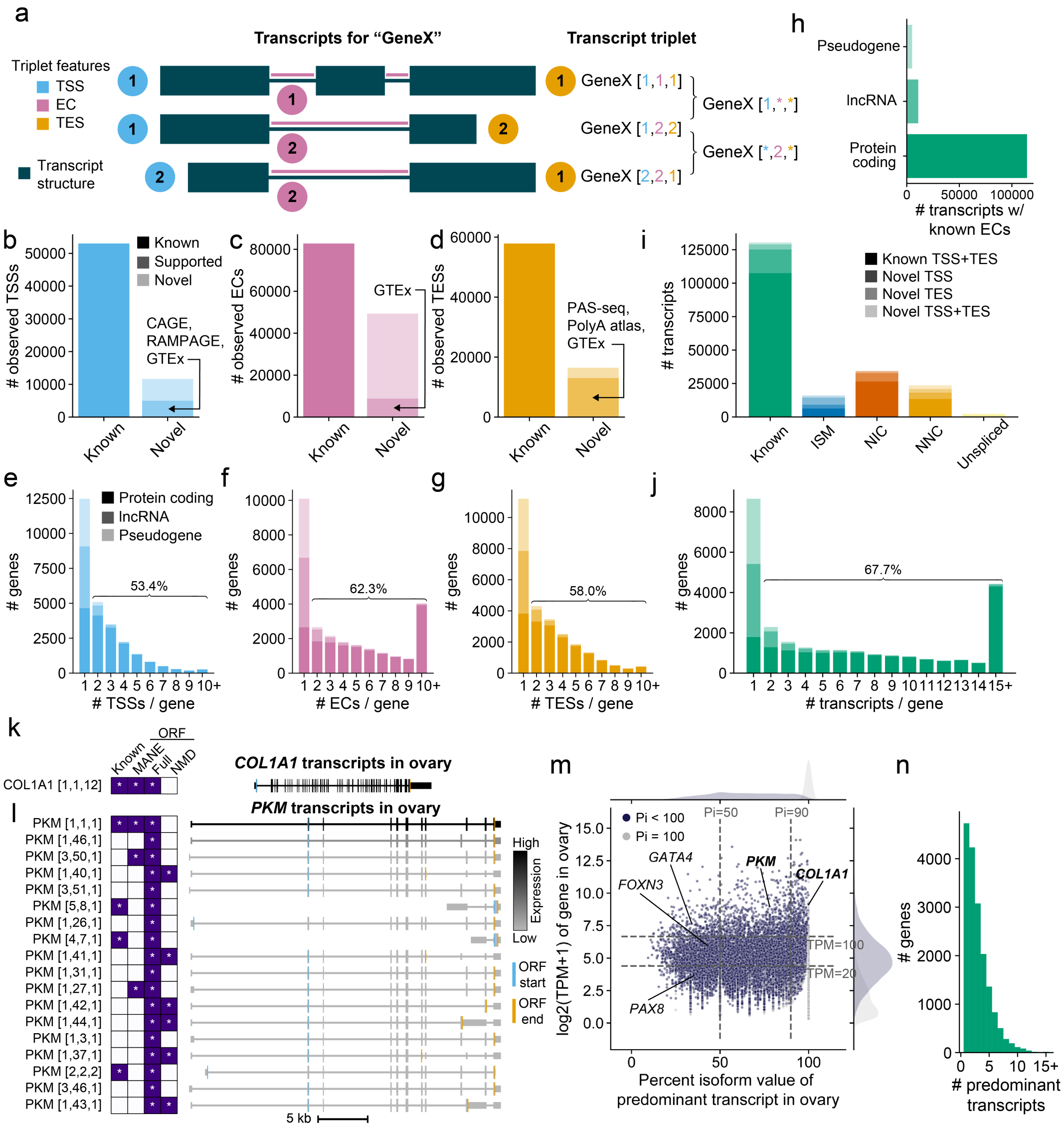


Figure 3.6: **Triplet annotation of transcript structure maps diversity within and across samples. a,** Transcript triplet naming convention for 3 transcripts from the same gene based on the transcript start site (TSS), exon junction chain (EC), and transcript end site (TES) used. **b-d,** Triplet features detected in human ENCODE LR-RNA-seq from GENCODE v40 polyA genes by novelty and support. Known features are annotated in GENCODE v29 or v40. **e-g,** Triplet features detected in human ENCODE LR-RNA-seq per GENCODE v40 polyA gene split by gene biotype. **h,** Transcripts from GENCODE v40 polyA genes detected from human ENCODE LR-RNA-seq that have a known EC by biotype. **i,** Novelty characterization of triplet features in each transcript detected in the human ENCODE LR-RNA-seq. **j,** Transcripts detected in human ENCODE LR-RNA-seq per GENCODE v40 polyA gene by biotype. **k,** *COL1A1* transcripts expressed in the ovary sample from human ENCODE LR-RNA-seq. **l,** *PKM* transcripts expressed 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 (pi) value of the predominant transcript for each gene expressed in human ovary sample. **n,** Number of unique predominant transcript per gene.

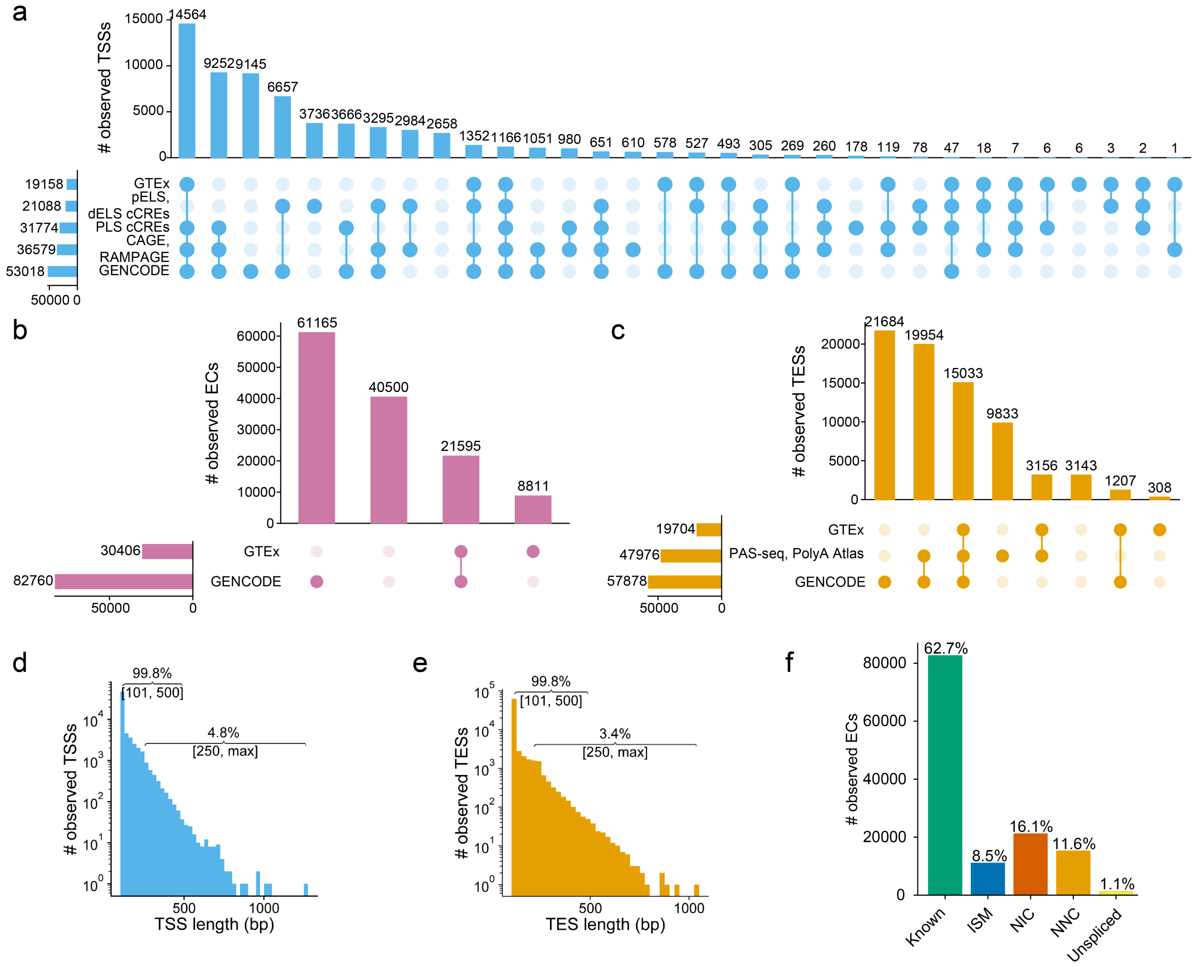


Figure 3.7: **Characterization of observed triplet features from human LR-RNA-seq. a-c,**Upset plots showing sources that overlap observed triplet features derived from human ENCODE LR-RNA-seq for **a,** TSSs **b,** ECs **c,** TESs. **d-e,** Lengths of observed **d,** TSSs **e,** TESs derived from human ENCODE LR-RNA-seq. **f** Novelty of unique ECs detected ≥ 1 TPM from polyA genes in human ENCODE LR-RNA-seq.

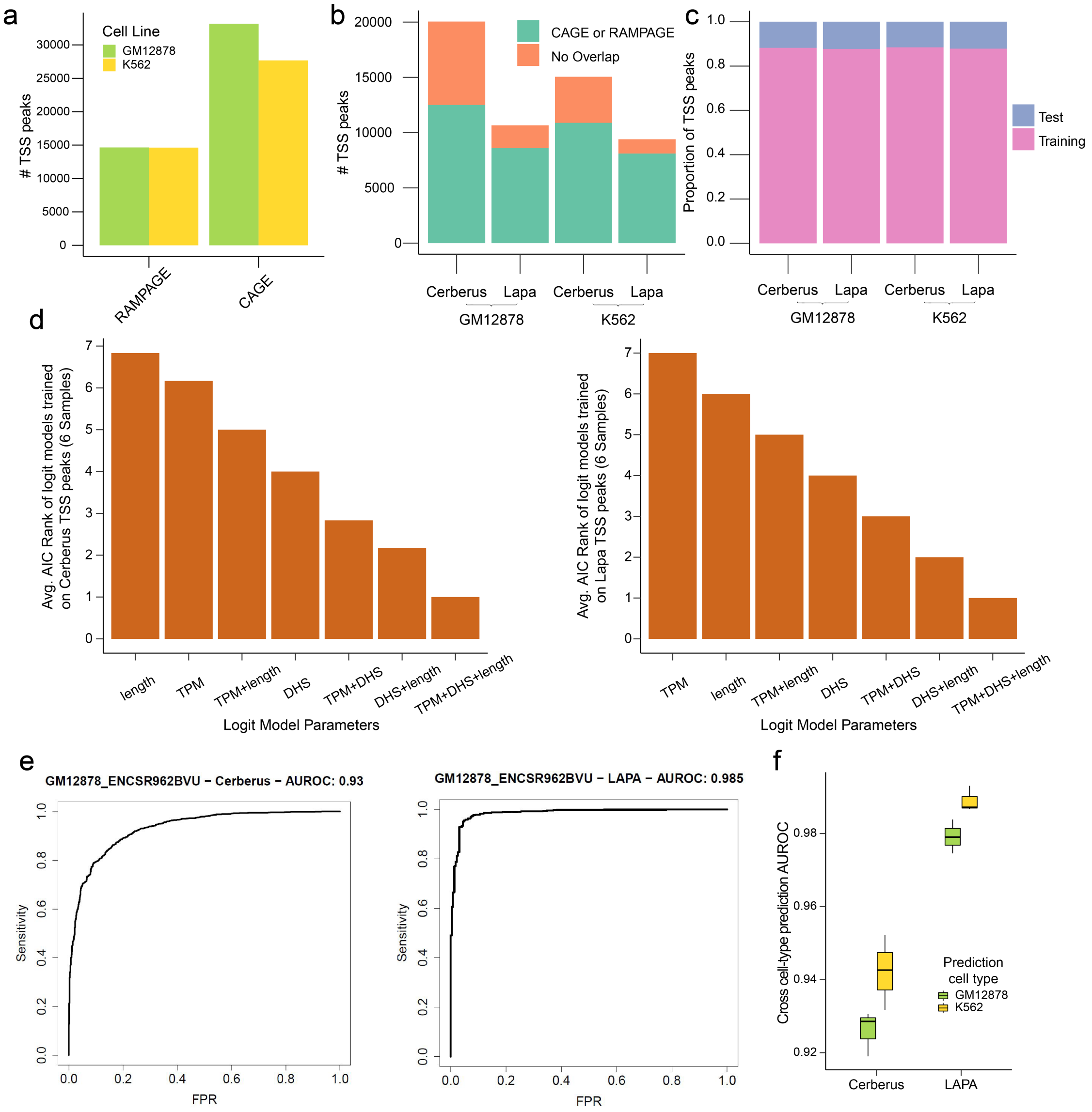


Figure 3.8: **Machine learning models predict support for long-read TSS peaks by other TSS assays in a cross-cell type manner. a,** Number of long-read RNA-seq TSS peaks called by RAMPAGE and CAGE. **b,** Number of long-read TSS peaks called by Cerberus or LAPA supported by RAMPAGE or CAGE assays in 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 K562 RAMPAGE and CAGE in a cross-cell type manner, a model is trained on a GM12878.

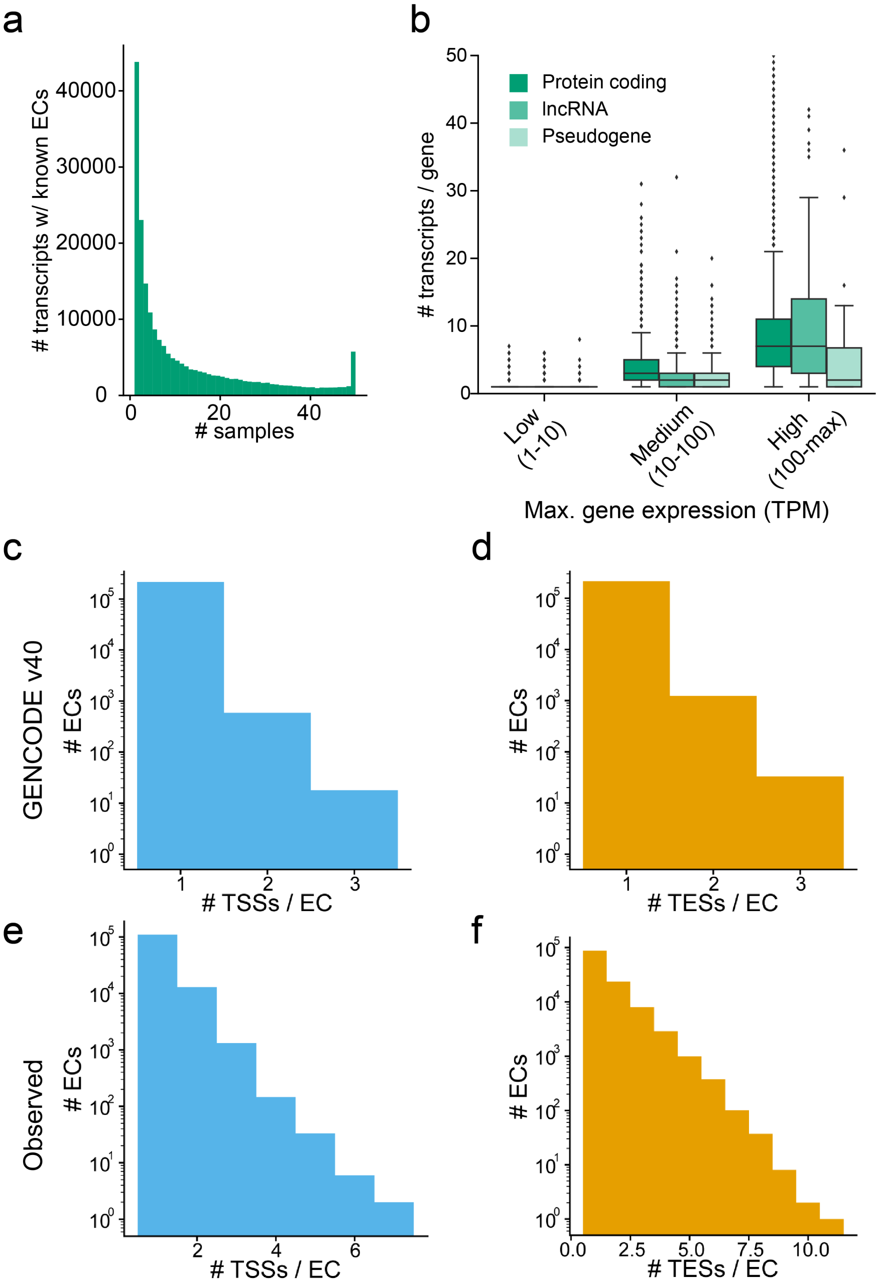
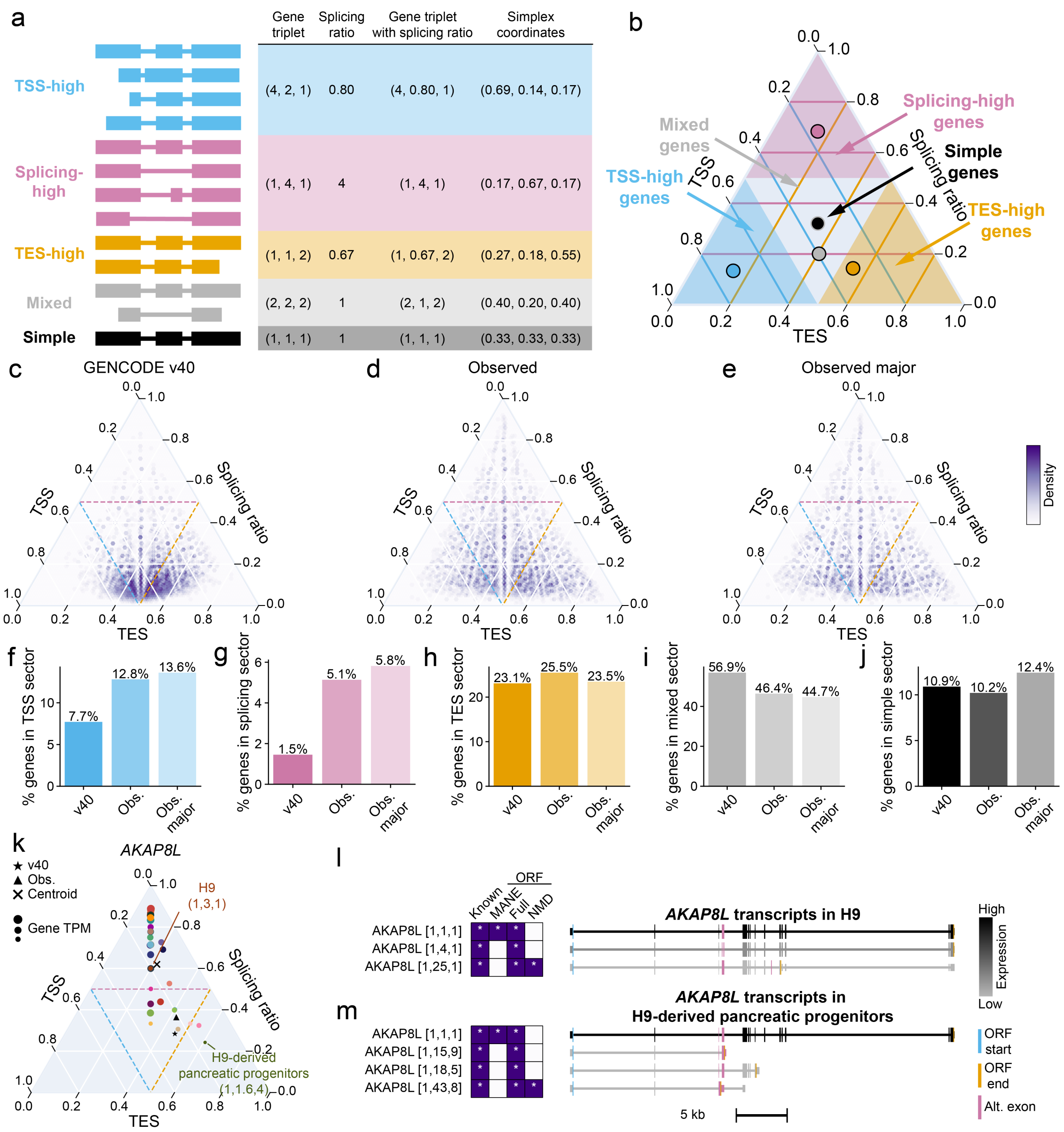


Figure 3.9: **Characterization of observed transcripts from human LR-RNA-seq.** **a,** Number of samples in which each transcript with a known EC is detected ≥ 1 TPM in the human ENCODE LR-RNA-seq dataset. **b,** Boxplot of, for the sample where a gene is most highly expressed, the number of transcripts 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.



	Gene triplet	Splicing ratio	Gene triplet with splicing ratio	Simplex coordinates
TSS-high	(4, 2, 1)	0.80	(4, 0.80, 1)	(0.69, 0.14, 0.17)
Splicing-high	(1, 4, 1)	4	(1, 4, 1)	(0.17, 0.67, 0.17)
TES-high	(1, 1, 2)	0.67	(1, 0.67, 2)	(0.27, 0.18, 0.55)
Mixed	(2, 2, 2)	1	(2, 1, 2)	(0.40, 0.20, 0.40)
Simple	(1, 1, 1)	1	(1, 1, 1)	(0.33, 0.33, 0.33)

			ORF	
	Known	MANE	Full	NMD
AKAP8L [1,1,1]	*	*	*	
AKAP8L [1,4,1]	*		*	
AKAP8L [1,25,1]	*		*	*

			ORF	
	Known	MANE	Full	NMD
AKAP8L [1,1,1]	*	*	*	
AKAP8L [1,15,9]	*		*	
AKAP8L [1,18,5]	*		*	
AKAP8L [1,43,8]	*		*	*

Figure 3.10: **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. **b,** Layout of the gene structure simplex with the genes from **a,** plotted based on their simplex coordinates. **c-e,** Gene structure simplices for the transcripts from protein coding genes that are **c,** annotated in GENCODE v40, **d,** the observed set of transcripts in human, **e,** the observed major set of transcripts, the union of major transcripts from each sample detected in human. **f-j,** Proportion of genes from the GENCODE v40, observed, and observed major sets that fall into the different sectors **k,** Gene structure simplex for *AKAP8L*. **l-m,** Transcripts of *AKAP8L* expressed ≥ 1 TPM in **l,** H9 **m,** H9-derived pancreatic progenitors colored by expression level in TPM.

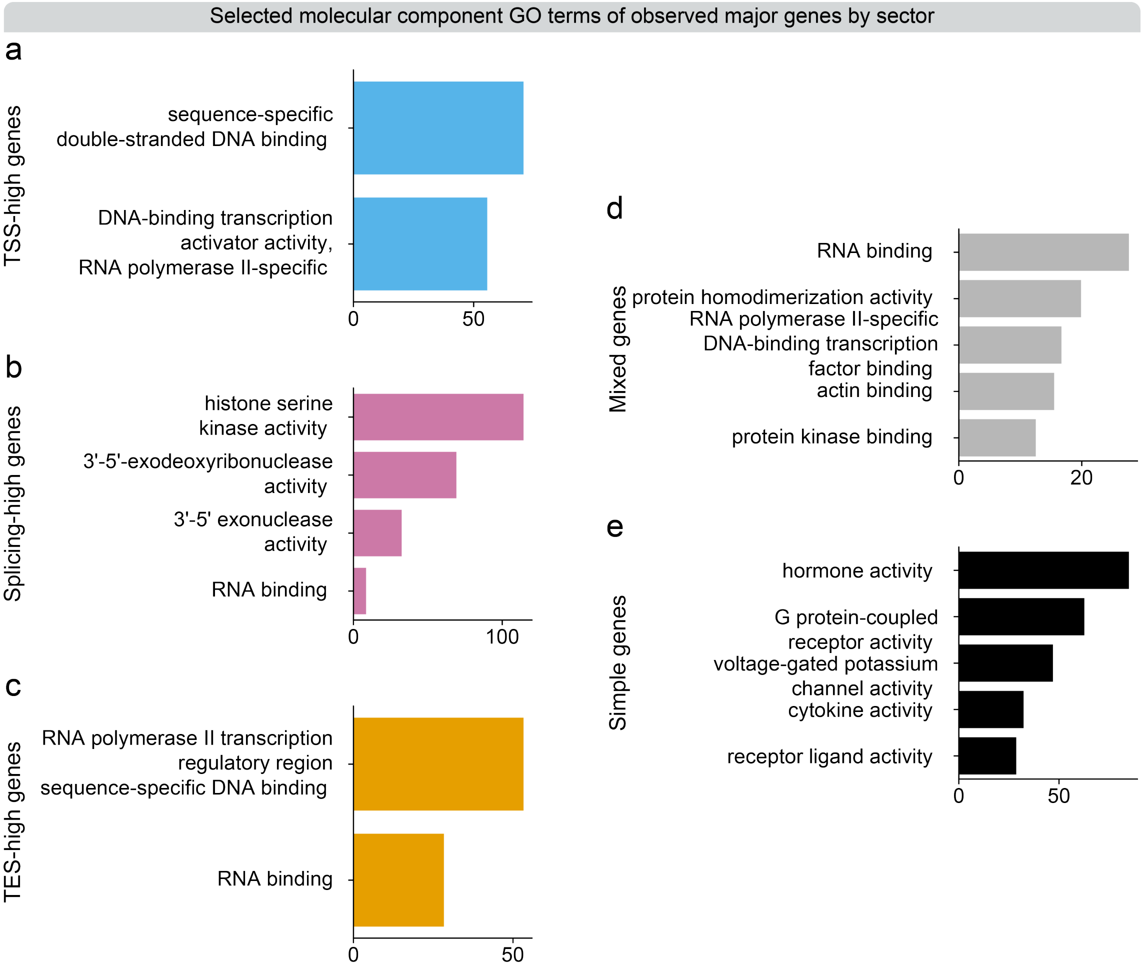


Figure 3.11: **Genes in different sectors are enriched for distinct molecular activities. a-e,** Selected GO terms from the top 5 results for observed major genes in sector for the **a,** TSS-high **b,** splicing-high **c,** TES-high **d**, mixed and **e,** simple sectors. X axis indicates Enrichr combined scores.

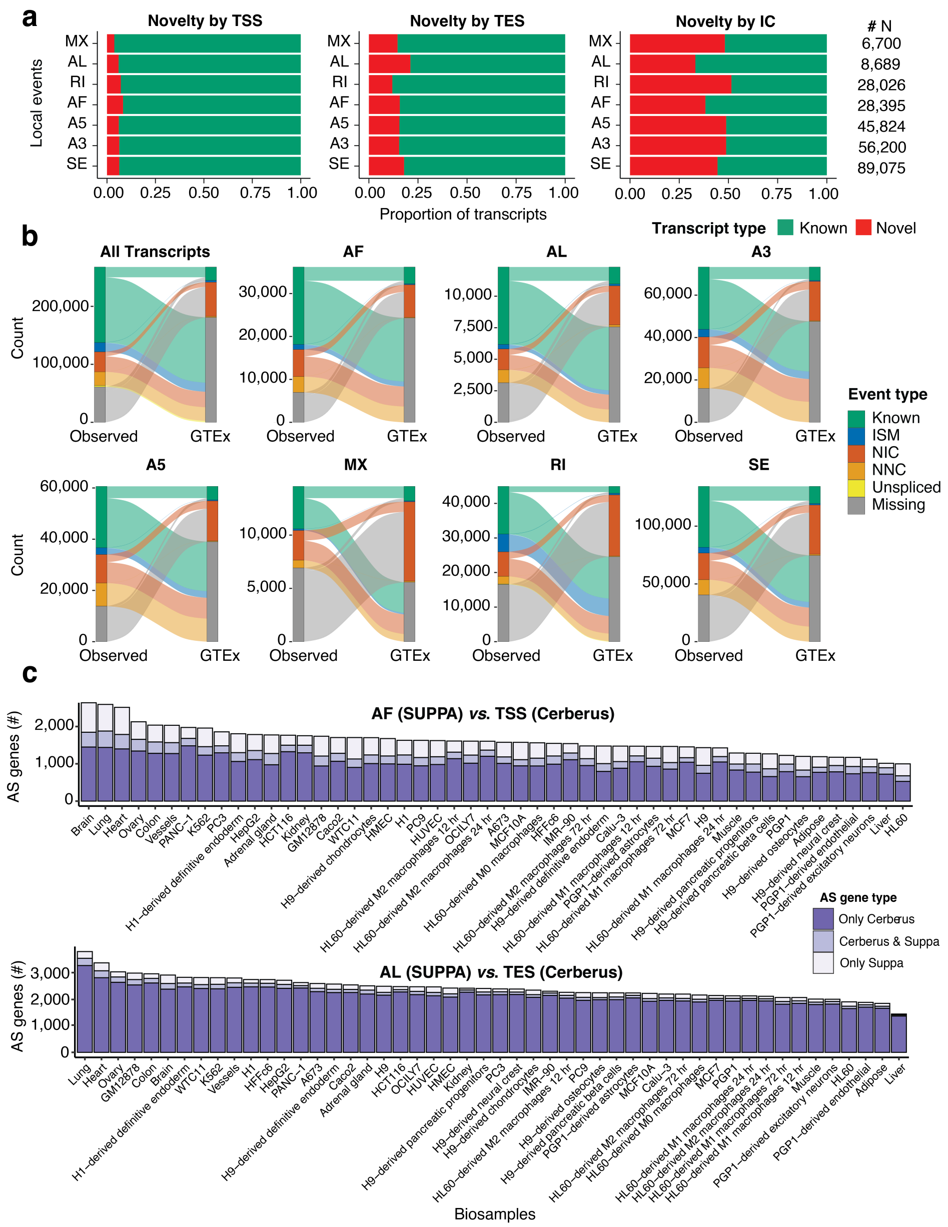
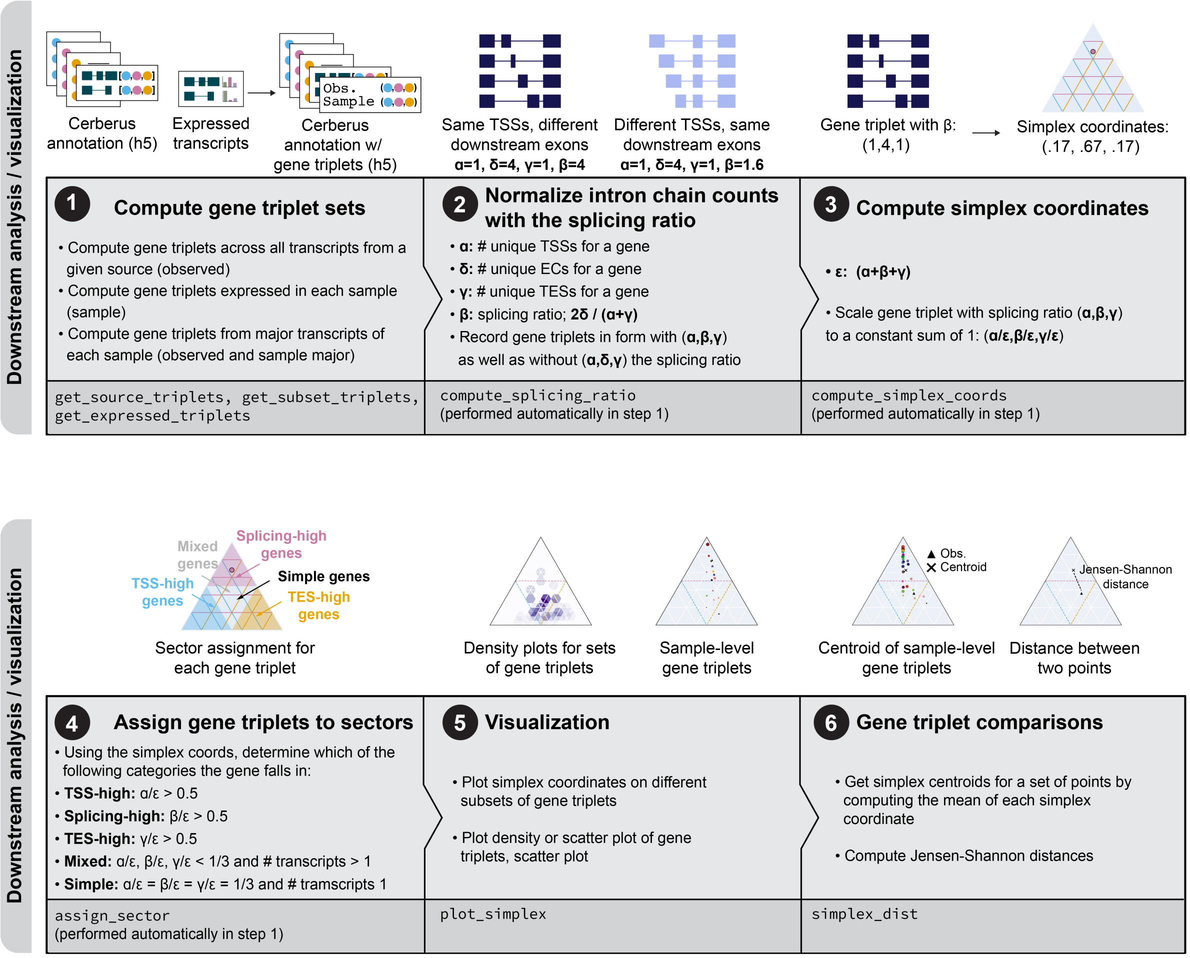


Figure 3.12: **Alternative splicing (AS) detection by SUPPA and Cerberus. a,** Barplots showing, for each type of local AS event detected by SUPPA (y axis), the proportion of observed known and novel transcripts identified by Cerberus (x axis), based on novel events at the TSS, TES, or EC. **b,** Sankey plots showing the number of transcripts classified as Known, Novel In Catalog (NIC), Novel Not in Catalog (NNC), Unspliced, or Missing by Cerberus and GTEx. 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.


Downstream analysis / visualization
Cerberus annotation (h5)
Expressed transcripts
Obs.
Sample
Cerberus annotation w/ gene triplets (h5)
Same TSSs, different downstream exons
α=1, δ=4, γ=1, β=4
Different TSSs, same downstream exons
α=1, δ=4, γ=1, β=1.6
Gene triplet with β: (1,4,1)
Simplex coordinates: (.17, .67, .17)

1 Compute gene triplet sets

- Compute gene triplets across all transcripts from a given source (observed)
- Compute gene triplets expressed in each sample (sample)
- Compute gene triplets from major transcripts of each sample (observed and sample major)

```
get_source_triplets, get_subset_triplets,
get_expressed_triplets
```

2 Normalize intron chain counts with the splicing ratio

- **α:** # unique TSSs for a gene
- **δ:** # unique ECs for a gene
- **γ:** # unique TESs for a gene
- **β:** splicing ratio; **2δ / ($\alpha+\gamma$)**
- Record gene triplets in form with **(α,β,γ)** as well as without **(α,δ,γ)** the splicing ratio

`compute_splicing_ratio`
(performed automatically in step 1)

3 Compute simplex coordinates

- **ε: ($\alpha+\beta+\gamma$)**
- Scale gene triplet with splicing ratio **(α,β,γ)** to a constant sum of 1: **($\alpha/\varepsilon,\beta/\varepsilon,\gamma/\varepsilon$)**

`compute_simplex_coords`
(performed automatically in step 1)

Downstream analysis / visualization
Mixed genes
Splicing-high genes
TSS-high genes
Simple genes
TES-high genes
Sector assignment for each gene triplet
Density plots for sets of gene triplets
Sample-level gene triplets
▲ Obs.
✕ Centroid
Centroid of sample-level gene triplets
Jensen-Shannon distance
Distance between two points

4 Assign gene triplets to sectors

- Using the simplex coords, determine which of the following categories the gene falls in:
- **TSS-high:** $\alpha/\varepsilon > 0.5$
- **Splicing-high:** $\beta/\varepsilon > 0.5$
- **TES-high:** $\gamma/\varepsilon > 0.5$
- **Mixed:** α/ε, β/ε, $\gamma/\varepsilon < 1/3$ and # transcripts > 1
- **Simple:** $\alpha/\varepsilon = \beta/\varepsilon = \gamma/\varepsilon = 1/3$ and # tramscripts 1

`assign_sector`
(performed automatically in step 1)

5 Visualization

- Plot simplex coordinates on different subsets of gene triplets
- Plot density or scatter plot of gene triplets, scatter plot

`plot_simplex`

6 Gene triplet comparisons

- Get simplex centroids for a set of points by computing the mean of each simplex coordinate
- Compute Jensen-Shannon distances

`simplex_dist`

Figure 3.13: **Overview of gene triplet based downstream analysis and visualization with Cerberus.**

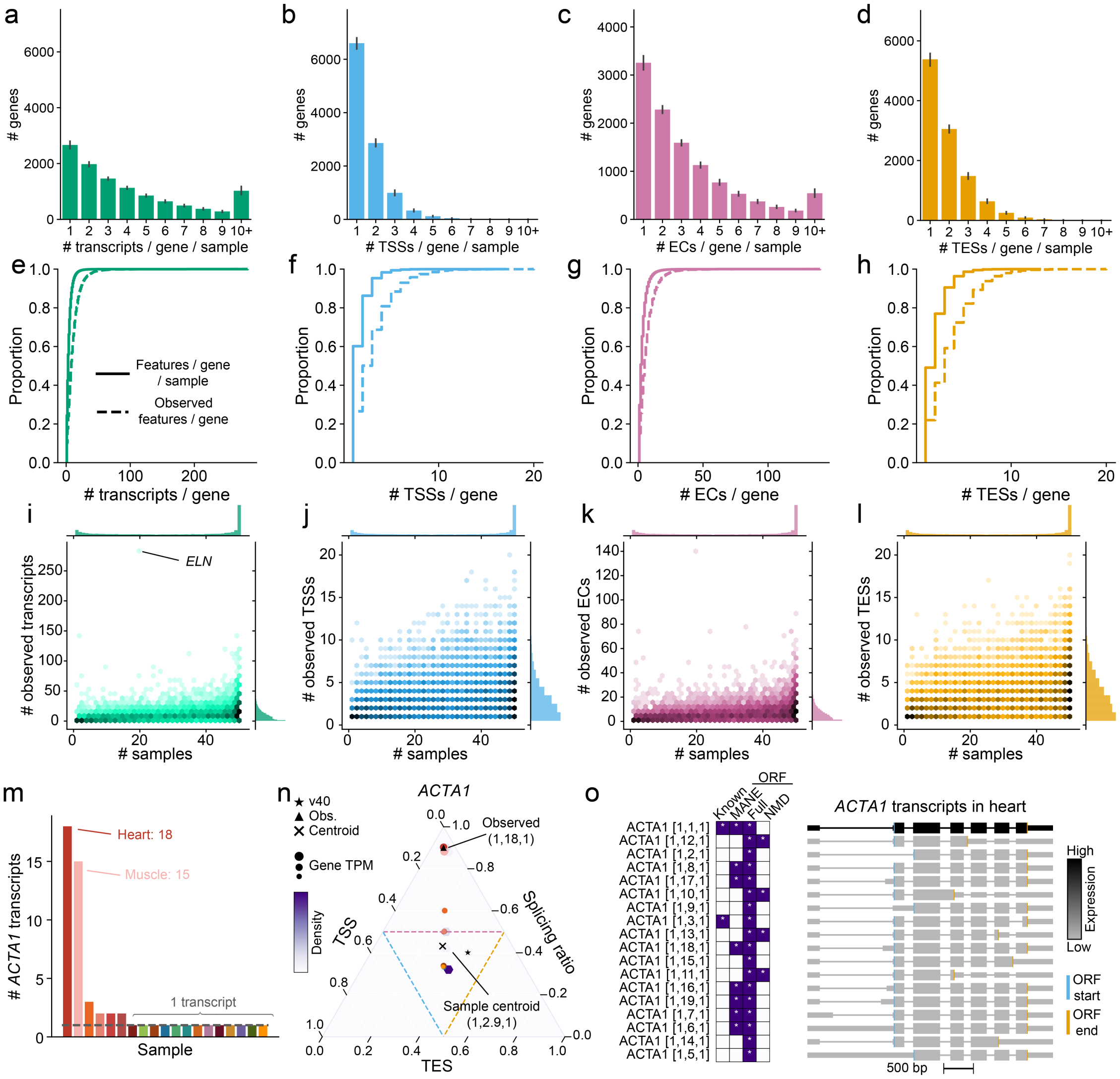


Figure 3.14: **Uncovering sample-specific behavior of triplet features by comparing observed and sample-level gene triplets. a-d,** Number of features detected per gene per sample for **a,** transcripts **b,** TSSs **c,** ECs **d,** TESs. **e-h,** Number of features per gene per sample and observed overall showing the proportion of the distribution that comes from each number of **e,** transcripts **f,** TSSs **g,** ECs **h,** TESs. **i-l,** Number of observed overall triplet features or transcripts per gene versus the number of samples each gene is expressed in for **i,** transcripts **j,** TSSs **k,** ECs **l,** TESs. **m,** Number of *ACTA1* transcripts expressed in each sample. **n,** Gene structure simplex for *ACTA1*. Gene triplets with splicing ratio for the overall observed and sample-level centroid labeled. Simplex coordinates for the GENCODE v40, observed set, and centroid of the samples also shown for *ACTA1*. **o,** Browser models of transcripts of *ACTA1* expressed ≥ 1 TPM in heart colored by expression level in TPM.

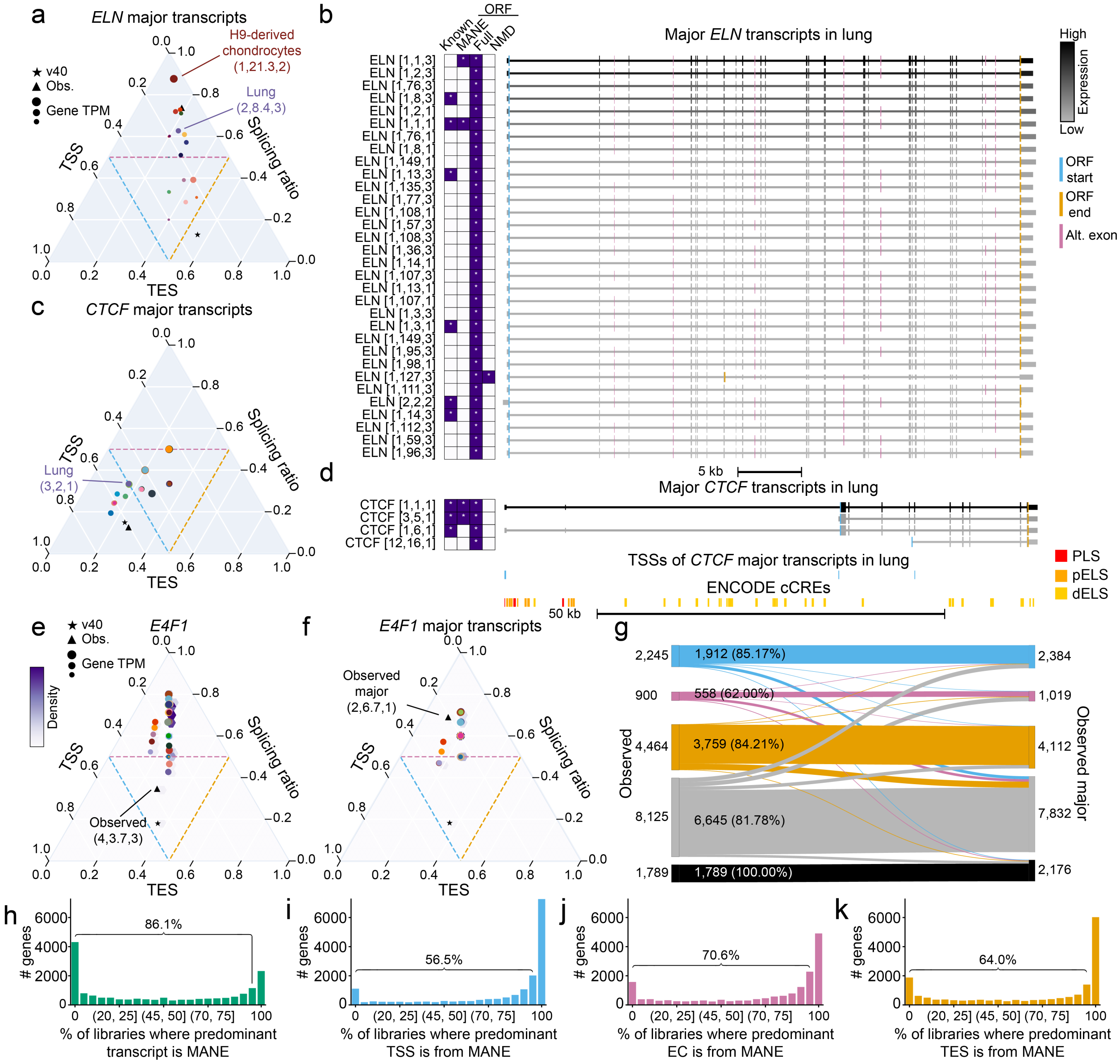


Figure 3.15: **Sample-specific and global changes in predominant and major transcript isoform usage.** **a,** Gene structure simplex for major transcripts of *ELN*. **b,** Major transcripts of *ELN* expressed ≥ 1 TPM in lung colored by expression level in TPM. **c,** Gene structure simplex for major transcripts of *CTCF*. **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*. **f,** Gene structure simplex for major transcripts of *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.

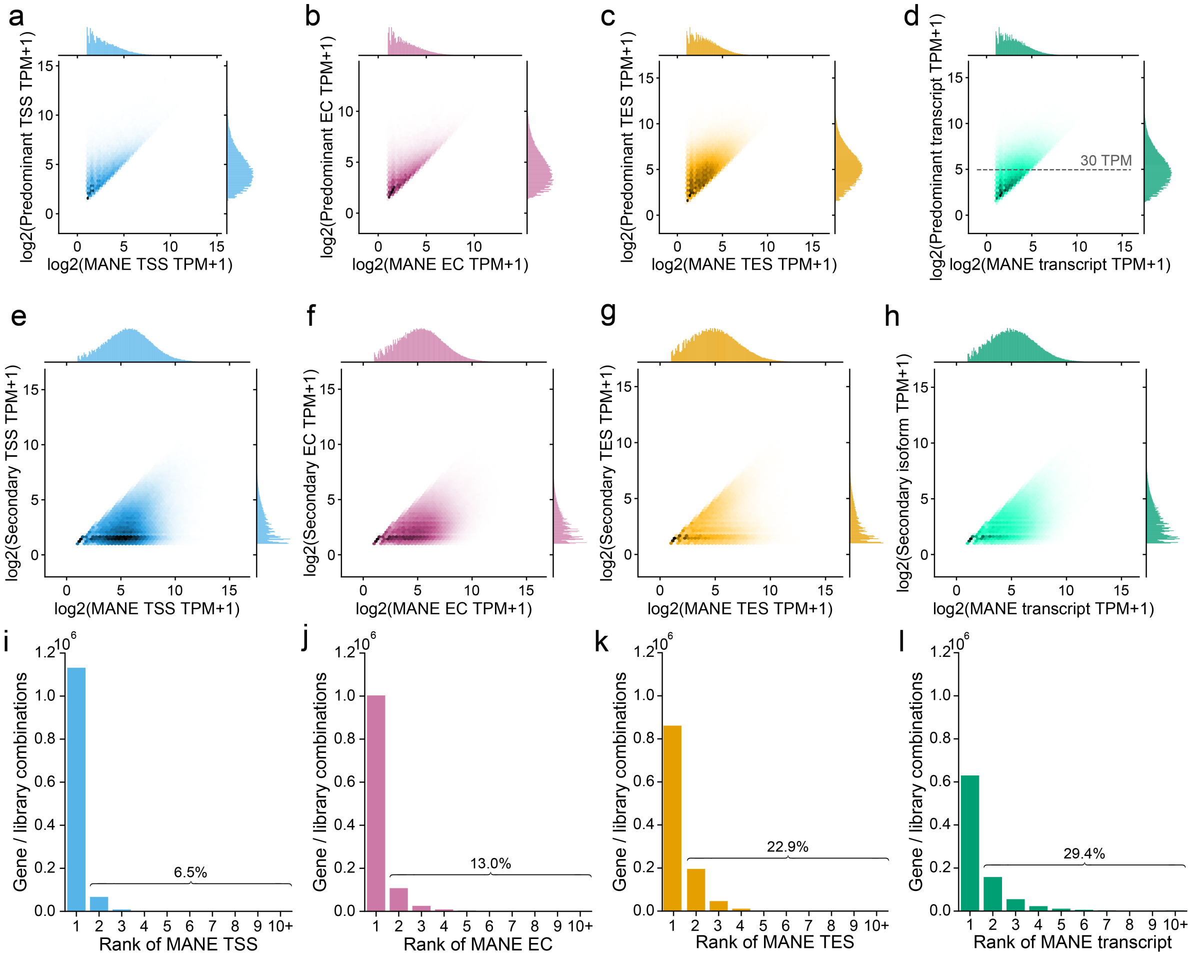


Figure 3.16: **Rank and expression of predominant and MANE transcripts. a-d,** For protein coding genes with MANE transcripts from GENCODE v40 where the predominant transcript or triplet feature is not the MANE transcript or triplet feature but is still expressed, expression of the predominant vs. the 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.

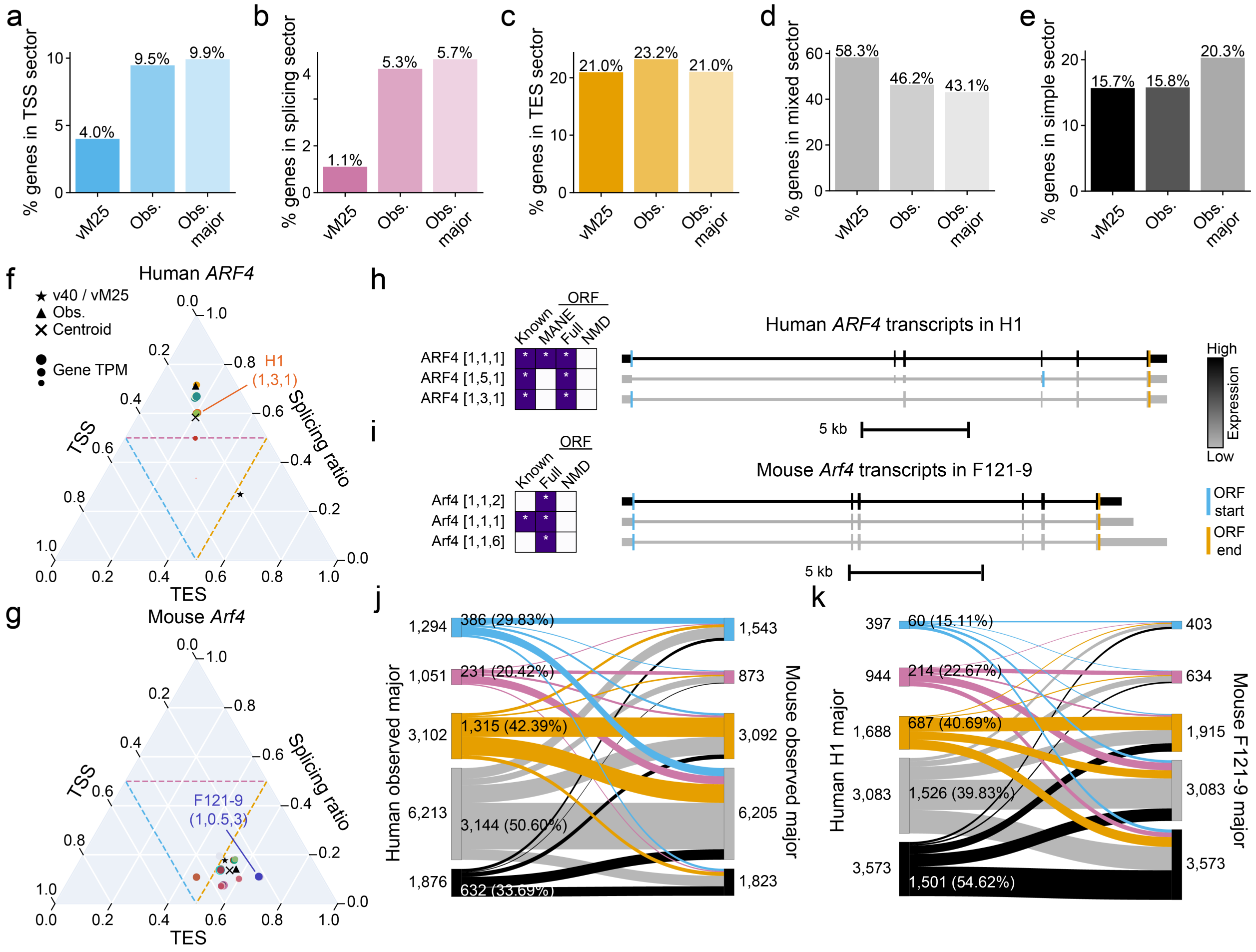


Figure 3.17: **Conservation of gene triplets from human and mouse. a-e,** Proportion of genes from the GENCODE vM25, observed, and observed major sets that fall into the **a,** TSS-high sector, **b,** splicing-high sector, **c,** TES-high sector, **d,** mixed sector, **e,** simple sector. **f,** Gene structure simplex for *ARF4* in human. Gene triplet with splicing ratio for *ARF4* transcripts in H1 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 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.

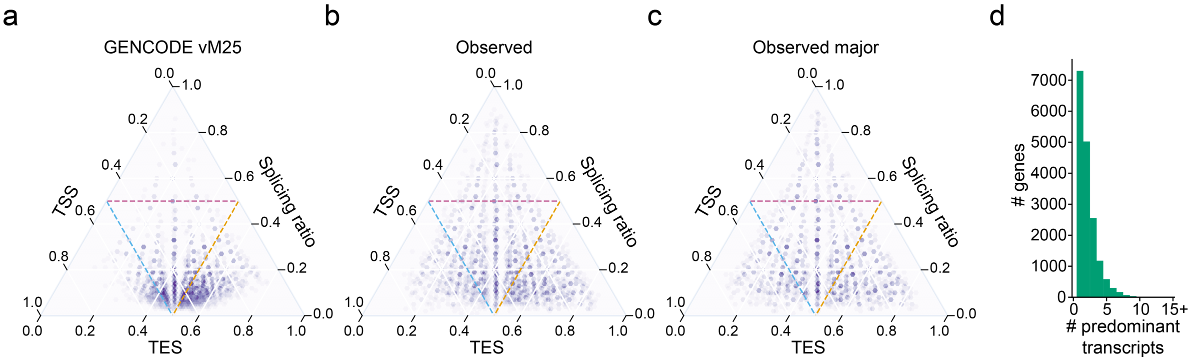


Figure 3.18: **The gene structure simplex for mouse. a-c,** Gene structure simplices for the transcripts from protein coding genes that are **a,** annotated in GENCODE vM25 where the parent gene is also detected in our mouse LR-RNA-seq dataset, **b,** the observed set of transcripts, those detected ≥ 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.

Chapter 4

Single-nucleus long-read RNA-seq of genetically distinct mouse strains reveals cell type and genotype-specific transcript usage

4.1 Abstract

Genetic variation and cell type-specific splicing are both contributors to alternative transcript usage. Here, we apply single-nucleus long-read RNA-seq (LR-Split-seq) to the prefrontal cortex of two genetically distinct mouse strains: C57B6/J and CAST/EiJ in order to evaluate the impact of both. We show that by using exome capture techniques, we are able to overcome the experimental limitations of sampling full-length RNA in the nucleus. We find that the genetic variation between strains is responsible for driving differential transcript, transcript start site (TSS), and transcript end site (TES) usage in distinct cell types; and

that these differences are seen more frequently in neuronal than non-neuronal cell types. We demonstrate that the effects of genotype on transcript choice are pervasive and are often captured globally in addition to in individual cell types.

4.2 Introduction

Genetic variation profoundly influences alternative transcript usage, which can arise from 5' and 3' end choice, as well as alternative splicing[136,137]. Alternative transcript usage can be intrinsic and required for defining a specific cellular program, and disruption of appropriate transcript expression can be pathogenic. An estimated that 15-60% of all identified disease-associated single nucleotide variants are related to splicing[13,14]. Furthermore, cancer, a disease of altered genetics, is highly enriched for splicing alterations[138] and thus demonstrates the impact of genetic variation or changes on transcript choice. The IGVF Consortium[139] seeks to identify the impact of genetic variation on function, including splicing using a variety of techniques, such as single-nucleus long-read sequencing[96].

The brain harbors some of the most complex splicing and transcript structure changes among human tissues[140,141]. Microexons, which are defined as exons less than < 27 bp in length, exhibit complex coordinated inclusion and exclusion events that frequently modify underlying protein function in the brain. These events show particularly high variability among neuronal cell types and are highly conserved between human and mouse[76,142,143]. Furthermore, misregulation of microoexon coordination is enriched in neurodevelopmental disorders[142]. 3' UTRs are typically longer in neurons compared to non-neuronal cell types, leading to high 3' end diversity of transcripts in the brain deriving from the heterogeneous cell type composition of neurons and non-neurons in the brain[144]. Prior single-cell transcript-resolved efforts have investigated how the apparent transcript diversity in the brain arises and found that changes in transcript usage in mouse are governed by cell type differences rather than regional differences in the brain, highlighting the importance of resolving transcript usage based on cell type rather than bulk brain region[57]. Furthermore, drivers of differential transcript isoform usage differ between cell types in the brain, where some are more variable based on subtype, brain region, or developmental age[94].

Given the extensive role of genetic variation in influencing transcript isoform choice, the abundance of alternative splicing and transcript usage in the brain, and the impact of cell type on transcript usage, it is important to determine to what extent changes in transcriptional outcomes are driven by change in cell type versus change in genotype in the brain. Though cell type-resolved studies have in the past relied on FACS for isolation of specific cell populations, single cell and single nucleus techniques are capable of sequencing the RNA present in individual cells or nuclei. Single nucleus experiments present a distinct problem when applied to full-length transcriptomics. The nucleus is the site of RNA transcription and splicingand therefore many sequenced molecules are unspliced[96]. While still indicative of expression on the gene level, which is the focus when performing traditional short-read single-nucleus RNA-seq (snRNA-seq), unspliced reads are mostlynearly entirely uninformative for long-read snRNA-seq data, where the goal is to quantify fully-processed, spliced transcript isoforms.

Here, we apply combinatorial barcoding-based long-read snRNA-seq (LR-Split-seq) and short-read snRNA-seq to the same nuclei from the left cerebral cortex of two divergent inbred mouse strains: C57B6/J and CAST/EiJ, which are part of the IGVF bridge sample collection. We show that using targeted exome capture highly enriches for spliced transcripts during single nucleus preparation though it does uniformly bias the gene expression profile toward genes with less total intron length. We characterize global and cell type-specific differences in transcript isoform usage across the main cell types in the brain, which are particularly pronounced in neuronal cell types. We illustrate that the impact of genotype on transcript choice is widespread, often evident both globally and within individual cell types, which contrasts with our findings when comparing sexes where we call proportionally more differential transcript usage events in specific cell types.

4.3 Results

Targeted exome capture increases spliced RNA yield from single nuclei

We performed LR-Split-seq[96] on left cortices of two distinct mouse genotypes, C57BL6/J and CAST/EiJ, with 4 biological replicates each (2 males and 2 females per genotype) all at 10 weeks of age (Fig. 4.1a). We sequenced two subpools of 13,000 nuclei using both short read (Illumina NextSeq2000) and long read sequencing using two Oxford Nanopore (ONT) platforms, which enables us to easily extend annotations or conclusions from the short read to the long read and vice versa.

Previous studies applying LR-RNA-seq to single nuclei have demonstrated substantial enrichment of unspliced, premature mRNA reads compared to full-length mature mRNAs[96]. We used a targeted exome panel which should probe for exon-containing molecules to enrich for fully-processed mRNA in single nuclei (Fig. 4.1a, Table 4.1). For the long read component, we first produced 60 million ONT MinION reads total for each of the exome and non-exome capture subpools. There was no effective difference in read lengths from each technique (Fig. 4.1b). However, we saw a significant reduction of the percentage of monoexonic reads in the exome capture (two-sided KS test statistic = 0.44; p-value=0.0) (Fig. 4.1c). Furthermore, when comparing all demultiplexed reads that were subject to exome capture versus those that were not, we observed a 2-fold reduction in the number of genomic reads, which are monoexonic and often intron-overlapping (Fig. 1.2, Fig. 4.1d-e). Thus our exome capture reads help us overcome the limitations of RNA sampling in the nucleus and provide us with more informative reads to study full-length, spliced transcript isoform usage

Unbiased clustering showed clear differences between transcriptome signatures between exome capture and non-exome capture data (Fig. 4.2a-c). Given the clear separation, we sought to characterize differentially expressed genes between the two techniques. We performed differential gene expression on our LR-Split-seq data between the exome capture and non-exome capture subpools. We found that the 4,942 genes upregulated in the non-exome capture have a significantly longer median total intron length than those upregulated in exome capture (Two-sided KS test statistic: 0.54, p-val: 2.4e-06) (Fig. 4.2d-e). This is likely due to genes with a longer intron length providing unprocessed pre-mRNAs with more potential sites to prime off of that would be filtered out with the exome capture.

LR-Split-seq recovers known cortical cell types

Given our success with the exome capture, we sequenced 106 million ONT PromethION reads on the same exome capture subpool. We annotated cell types in the cortex using the short read Split-seq across all of the subpools, including the exome capture and non exome capture (Methods, Fig. 4.3a-c). Using the annotations from matched nuclei between the short and long read, we found populations of canonical brain-resident cell types; dominated mainly by the glutamatergic neurons (GLUT), GABAergic neurons (GABA), and astrocytes (n = 9,170 total nuclei) (Table 4.2, Fig. 4.4a-d). As we relied on having matching nuclei from the short and long read to call cell types, we discarded any nuclei that were absent from or did not pass QC in the short-read Split-seq (Methods). As validation, we looked for marker gene expression of known cell types in the brain and found strong cell type-specific expression of said markers (Fig. 4.4e-f). We further recovered one versus all marker genes distinguishing each cell type, many of which are corroborated by marker genes from the literature (Fig. 4.4e-f, Fig. 4.5a).

On the transcript level, in order to be conservative in calling novel transcripts, we filtered out novel transcripts that were not previously identified by consortium-level bulk LR-RNA-seq (Methods)[101]. After filtering, we were left with 8,758 nuclei (Methods). We called marker transcripts for each cell type, some of which are transcripts of canonical marker genes, such as *Slc17a7*, *Apoe*, *P2ry12*, and *Cst2* (Fig. 4.5b). We investigated how often different transcripts from the same gene are called as markers across the cell types. Limiting marker transcripts to the top reported one for each gene and cell type, we found that 18 genes used different marker transcripts for different cell types. One such example was *Vamp2*, which encodes for a protein that is a part of the SNARE complex which is responsible for neurotransmitter release[145] (Fig. 4.5c-d).

Inter-cell type transcript isoform usage differs mainly in neuronal cell types

One of the current goals of single cell or nucleus LR-RNA-seq is to identify key transcript isoform usage difference between distinct cell types. Hence, we conducted differential usage (DU) testing, a method that detects genes involved in transcript isoform switching by examining alterations in transcript expression in relation to their parent gene[55,57,146] between each pair of cell types using Swan[56]. Our framework of characterizing transcripts based on their transcript start site (TSS), exon junction chain (EC), and transcript end site (TES) used allows for us to not only perform this DU testing on the full-length transcript isoform level (DU-T), but also on the TSS (DU-TSS), EC (DU-EC), and TES (DU-TES) level (Fig. 3.6a)[101]. We found the most striking differences between the neuronal cell types and astrocytes (168 for GABAergic, 283 for glutamatergic), followed by differences between the neuronal cell types and oligodendrocytes and oligodendrocyte precursors (118 between oligodendrocytes and glutamatergic neurons) (Fig. 4.6a-b, Fig. 4.7). Overall, both neuronal cell types showed the most DU genes total across all cell types (Fig. 4.6b, Fig. 4.7c-f). These results indicate that transcript isoform usage is the most different in neurons compared to the rest of the brain-resident cell types. When considering all forms of differential transcript feature usage, we find that the main differences are in the full-length isoform and exon junction chain (32.2% and 30.7% of all called DU respectively) (Fig. 4.6c). Remarkably, between different cell types, the proportionally fewest events were called at the TSS level (16.1%)

(Fig. 4.6c). For example, we found *Cadm1*, which is a cell adhesion molecule gene, was found to to be DU-T / DU-EC between both GABAergic and glutamatergic neurons versus the astrocytes as previously reported in studies using single-nucleus long-read RNA-seq[76,147] (Fig. 4.6d). Remarkably, both the neuronal and astrocyte transcript isoforms use the same TSS and TES but use a different EC which is characterized by an exon skipping event in neurons.

Given our higher nuclei count and read count for GABAergic and glutamatergic neurons, we wanted to characterize inter-neuronal variability of transcript usage at the sub type level. Again using annotations that were performed in the corresponding short read nuclei, we annotated neuronal subtypes and thresholded the subtypes used based on excitatory neuron subtypes that had enough reads for testing (Methods, Table 4.3). We found the most DU-T genes between the L2/3 intratelencephalic neurons and the other layers, with the highest reported DU-T genes between them and the L5 pyramidal tract neurons (n = 30) (Fig. 4.8a, Fig. 4.9a-c). Interestingly, the second highest number of DU-T genes was between the L2/3 and L2 intratelencephalic neurons (n = 19) which demonstrates that there is a clear distinction in transcriptional signature between intratelencephalic neurons in the different cortical layers despite their spatial proximity (Fig. 4.8a). Overall, the L2/3 intratelencephalic neurons demonstrated the most differences at the level of DU across the excitatory neuronal subtypes (Fig. 4.8b, Fig. 4.9d-f). Given the range of read depths associated with each excitatory neuron subtype, we compared the number of DU-T genes we called for each subtype to the number of reads in said subtype. Though unsurprisingly the highest number of DU-T events we called was in the L2/3 intratelencephalic subtype with the highest number of reads, an increase in read number did not correspond to a corresponding increase in events for the other subtypes (Fig. 4.3c). When comparing our number of DU events called across each transcript feature, we found once again that they are dominated by DU-T and DU-EC (27.3% and 28.8% respectively), with relatively fewer contributions from DU-TSS and DU-TES (Fig. 4.8d).

Currently, the specificity of TSS usage between cell types has not been thoroughly investigated in a single cell level, though it is generally agreed upon that TSSs are largely context-specific[74,148,149]. Given this, we investigated to what extent DU-TSS events drive differences in tissue identity versus cell type identity. Using the bulk ENCODE4 mouse LR-RNA-seq dataset[101], we performed pairwise DU tests for all transcript features between five tissues at postnatal month 2: hippocampus, cortex, adrenal gland, and skeletal muscle (Methods). We found that overall, the fraction of DU-TSS events between tissues resembles the fraction of DU-TSS events between cell types, thus implying that TSS usage plays a similar role in defining tissue and cell type identity, rather than playing a more distinguished role in one or the other (Fig. 4.10).

Transcript diversity outlines to what extent multiple transcript isoforms are expressed in the same biological context. We used the gene triplet-based method of assessing transcript diversity in a cell type-specific manner to determine how coexpressed transcripts in the same cell type typically vary from one another[101]. We found that across the glutamatergic neuron population, genes are more likely to express more than one transcript than for astrocytes (45.9% of genes versus 39.1% of genes) (Fig. 4.11a-b). However, by comparing to specific glutamatergic neuron and astrocyte samples from the bulk human ENCODE4 LR-RNA-seq data, it appears that our we are underestimating the transcript diversity in these populations, where we see 60.2% and 67.9% of genes that express more than one transcript (Fig. 4.11c-d). It appears that we are particularly underestimating the contribution of alternative splicing and polyadenylation to transcript diversity, though notably the fraction of genes with their predominant mode of transcript diversity deriving from alternative TSS usage does not change nearly as much (Fig. 4.11).

Though the apparent diversity of genes in our LR-Split-seq data did not recapitulate the same levels of diversity observed in corresponding bulk cell types, we still wanted to assess what the function of genes with many observed transcript isoforms was in our distinct cell types. We took the top 100 genes with the most expressed transcript isoforms in each cell type and performed GO enrichment (Methods). Broadly, we found that transcriptionally-complex genes were involved in RNA splicing and processing, neuronal development, and synaptic function (Fig. 4.12). Enrichment for self-regulatory RNA processing mechanisms and neuronal function is a common feature of genes with a large number of transcript isoforms and changes[17].

LR-Split-seq enables identification of cell type-specific transcript isoform switching events between genotypes

We performed DU testing between our two genotypes in each cell type individually, again for each transcript feature (full-length transcript, TSS, EC, and TES). Overall, we found again that neuronal cell types have more DU events across all cell types between the genotypes (Fig. 4.13a-b, 4.14a-c). However, we cannot rule out the possibility that the increased number of DU-T genes in neurons is higher simply because of their increased abundance in the overall mixture of cell populations in the brain, and thus the larger number of reads we devote to sequencing them, which is a trend we do see for these comparisons (Fig. 4.13c). We find similar rates of DU genes called based on which transcript feature is considered to when we compare cell types or subtypes (Fig. 4.13d).

We wanted to assess whether the genes we called as DU between genotypes would have been called so across the entire dataset rather than in individual cell types. Agnostic of cell type, we also called DU genes for all transcript features across genotypes and intersected the events called at the cell type level with those called globally (Fig. 4.13a). We found that 17.0% of DU-T genes were exclusively found between genotypes in specific cell types, indicating that a subset of genotype-induced transcript isoform switches are only detectable in specific cell types, which the LR-Split-seq data enables us to identify (Fig. 4.13e, Fig. 4.14d-f). We investigated which cell types the cell type-specific genotype DU events were present in. We found that, consistent with our overall results, the largest proportion of these DU events occurred in neuronal cell types (Fig. 4.13f, Fig. 4.14g-i). Furthermore, only a handful of cell type-specific genotype DU events were present in more than one cell type, indicating the specificity of these transcript isoform switches (Fig. 4.13f, Fig. 4.14g-i). One such switch only detected between genotypes in a specific cell type was *Calm1* in GABAergic neurons, called by both DU-T and DU-EC tests. *Calm1* plays a crucial role in normal neural function and development. It is associated with two known 3' UTR isoforms, and maintaining the right balance between them is critical for neurodevelopment and brain function[150]. Here, we see a switch between a completely spliced, annotated version of the gene with the longer 3' UTR and a version of the gene that appears to have a retained intron and a shorter 3' UTR (Fig. 4.13g).

For DU-T genes between the two genotypes that were called on the cell type and genotype-wide level, we found several notable examples. *Plp1* encodes for a protein that is a critical component of the myelin sheath that oligodendrocytes produce[151]. The gene produces two known spliced transcript isoforms each with an associated protein (PLP for the longer and DM20 for the shorter), and the balance of these transcripts is modulated during development[151]. Between genotypes, *Plp1* is called as DU in oligodendrocytes based on the relative abundance of said transcript isoforms. The CAST/EiJ mice use relatively more of the shorter transcript isoform (DM20), which is the transcript isoform that typically dominates during early myelination[152], compared to the C57BL/6J mice (Fig. 4.15a). Additionally, we found that *Apoe*, which harbors the allele with the highest associated risk for developing Alzheimer's disease[153], was called as DIE between genotypes in the astrocytes specifically. In this switch, we see the higher relative prevalence of a transcript isoform with a longer 3' UTR that spans an intron in C57BL/6J versus CAST/EiJ (Fig. 4.15b).

Sex-specific transcript usage varies less than genotype-specific transcript within cell types

We performed DU tests for each transcript feature between sexes in each cell type (Fig.

4.16a). Overall, we found far fewer DU genes between sexes in each cell type than we did between genotypes in each celltype, with the cell types with the most DU genes again found in the neurons (Fig. 4.13b, 4.16b, Fig. 4.14a-c). As compared to the intra-cell type genotype comparison, we called proportionally more DU-TSS events between sexes than we did between genotypes in individual cell types (Fig. 4.16c). We investigated the extent to which the intra-cell type comparisons capture global sex-specific DU patterns rather than true cell type-specific DU. There were proportionally more cell type-specific sex DU genes than there were for the genotype comparison (66.7% vs. 17.0% respectively), and all the cell type-specific DU events were limited to exactly one cell type (Fig. 4.13e, Fig. 4.16d-e, Fig. 4.17d-i). This suggests that while the number of DU events is lower between sexes compared to between genotypes, these distinctions are more likely to manifest in particular cell types rather than stemming from a comprehensive alteration in transcript preference, which contrasts to the results from comparing genotypes. Though the sex comparisons yielded fewer overall numbers, we still found interesting instances of DU-T between the sexes. For instance, *Mobp*, another protein critical for the myelin sheath deposited on axons by oligodendrocytes[154], was called as DU-T, DU-EC, and DU-TES between the sexes only in the cell type-specific comparison in the oligodendrocytes. The predominant transcript isoform in females is missing the last exon that is seen in the predominant transcript isoform in males. Furthermore, annotated CDS sequences indicate that the stop codon used is different between the two transcripts (Fig. 4.16f).

4.4 Discussion

Here we present a cell type, genotype, and sex-specific map of differential transcript feature usage in the cortices of C57BL/6J and CAST/EiJ mice. First, we showed that using exome capture increased the yield of spliced RNA; decreasing the percentage of monoexonic reads from 80% to 40%, but that this capture strategy biases RNA sampling towards genes with lower total intron length.

Using our DU approach, we are able to capture important sources of variation in transcript, TSS, EC, and TES variation between genotypes, sexes, and cell types, and assess the extent to which each feature plays a role in defining the transcript usage between contexts. Overall, we found that DU events across all comparisons were dominant in the glutamatergic and GABAergic neurons, though we suspect in many cases this is due to low sampling depth in the other cell types, and reason that with increased read yield, we would uncover more differences in transcript feature usage.

By comparing transcript feature usage globally and in specific cell types between genotypes, we get a clearer picture of how often genotype drives global versus context-specific differences in transcript usage. Here, we found that the majority of DU events found in individual cell types were also called globally. This indicates that by and large, the effect of genetic variation influences transcript choice irrespective of cell type. However, there is still a subset of differences in transcript usage that are only reported in individual cell types, which highlights the importance of considering cell type-specific differences in addition.

We furthermore show that biological sex does not drive as striking of differences in transcript usage as genotype does. We detect far fewer DU events between sexes globally and in individual cell types. Interestingly though, we do not observe as striking of overlaps between global and cell type-specific switches between sexes as we do between genotypes. This implies that while there are fewer events overall, sex is more likely than genotype to drive differences in transcript usage within a specific cell type than globally.

This work represents a first effort towards the overarching goal of the IGVF consortium, which seeks to assess the impact of genetic variation on function[139]. As splicing and alternative transcript usage are heavily linked to genetic variation[136,137], an important aspect of understanding variant function is to determine the role they play in dictating transcriptional outcomes. This work highlights genotype comparisons at the global and cell type level and will serve as an analytical framework that can be harnessed for additional genotypes and tissue types in the future. Future IGVF work from our group will sequence additional inbred genetically-distinct mouse strains in eight different tissues using both short-read and long-read Split-seq. We estimate based on results here that transcript usage will vary mostly across cell types and genotypes, though some cell types will harbor more genetically-distinct DU signatures than others between genotypes. Based on the literature, we expect the most transcriptional diversity to be present in the brain and muscle tissues[140].

One limitation of this study is that our mapping was performed only using the C57BL/6J mm10 reference genome. Though clearly the correct reference for the C57BL/6J mice, our mapping for the CAST/EiJ mice was suboptimal based on its unique genome. As we extend this work to additional mouse genotypes, it will critical to map reads to their correct genotypes to minimize technical error that might arise from mapping issues, or to build a pangenome that represents the entire gamut of genetic diversity across the unique mouse strains.

4.5 Methods

Mice and tissue collection

Mice were housed at the UCI Transgenic Mouse Facility (TMF) in a temperature-controlled pathogen-free room under 12 hour light/dark cycles (lights on at 07:00 hr, off at 19:00 hr). The animal experiments were reviewed and approved by the Institutional Animal Care and Use Committee (IACUC), protocol AUP-21-106, "Mouse genomic variation at single cell resolution". Left cerebral cortex tissues of 10 week old mice were harvested from 12

C57BL/6J and 12 CAST/EiJ (2 males and 2 females per genotype) between the hours of 09:00 to 13:00. Tissues were stored in 1 mL Bambanker media in cryotubes kept at -80°C until nuclei isolation.

Purification of nuclei from mouse tissues

Tissues were thawed in Bambanker media on ice until the tissue could be extracted and lysed using Nuclei Extraction Buffer (Miltenyi Biotec cat. #130-128-024). Using forceps, tissues were transferred to a chilled gentle MACS C Tube (Miltenyi Biotec cat. #130-093-237) with 2 mL Nuclei Extraction Buffer supplemented with 0.2 U/uL RNase Inhibitor (New England Biolabs cat. M0314L). Nuclei were dissociated from whole tissue using a gentleMACS Octo Dissociator (Miltenyi Biotec cat. #130-095-937). The resulting suspension was filtered through a 70 um MACS SmartStrainer then a 30 um strainer (Miltenyi Biotec cat. 130-110-916 and #130-098-458, respectively). Nuclei were resuspended in 3 mL PBS + 7.5% BSA (Life Technologies cat. #15260037) and 0.2 U/ul RNase inhibitor for manual counting using a hemocytometer and DAPI stain (Thermo Fisher cat. #R37606).

Nuclei fixation

After counting, 4 million nuclei per sample were fixed using Parse Biosciences' Nuclei Fixation Kit v2 (cat. #ECF2003), following the manufacturer's protocol. Briefly, nuclei were incubated in fixation solution for 10 minutes on ice, followed by permeabilization for 3 minutes on ice. The reaction was quenched, then nuclei were centrifuged and resuspended in 300 uL Nuclei Buffer (Parse Biosciences cat. #ECF2003) for a final count. DMSO (Parse Biosciences cat. #ECF2003) was added before freezing fixed nuclei at -80°C in a Mr. Frosty (Sigma-Aldrich cat. #635639).

Split-seq experimental protocol

Nuclei were barcoded using Parse Biosciences' WT Kit v2 (cat. #ECW02030), following the manufacturer's protocol. Fixed, frozen nuclei were thawed in a 37°C water bath and added to the Round 1 reverse transcription barcoding plate at 19,500 nuclei per well, with alternating columns in rows A and C containing C57BL/6J males and females and rows B and D containing CAST/EiJ males and females. In situ reverse transcription (RT) and annealing of barcode 1 + linker was performed using a thermocycler (Bio-Rad T100, cat. #1861096). After RT, nuclei were pooled and distributed in 96 wells of the Round 2 ligation barcoding plate for the in situ barcode 2 + linker ligation. After Round 2 ligation, nuclei were pooled and redistributed into 96 wells of the Round 3 ligation barcoding plate for the in situ barcode 3 + UMI + Illumina adapter ligation. Finally, nuclei were counted using a hemocytometer and distributed into 8 subpools of 13,000 nuclei. The nuclei in each subpool were lysed and cDNA was purified using AMPure XP beads (Beckman Coulter cat. #A63881), then the barcoded cDNA underwent template switching and amplification. Importantly, for two subpools ("13G" and "13H") we increased the number of PCR cycles to 13 cycles from 12, and increased the extension time from 3 minutes to 13 minutes in order to increase the yield of full-length barcoded cDNA. cDNA from one of the subpools ("13G") also received exome capture treatment using Parse Biosciences' Custom Gene Capture Kit (cat. #GCE1001) and a Mouse Exome Panel (Twist Bioscience, cat. #102036). 1 ug of cDNA was hybridized with a blocker solution to block repetitive sequences, then hybridized with the exome panel overnight. Captured molecules were purified using Streptavidin beads, then amplified again using the cDNA amplification reagents from the WT Kit v2 (Parse Biosciences cat. #ECW02030). The cDNA for all 8 subpools were cleaned using AMPure XP beads and quality checked using an Agilent Bioanalyzer before proceeding to Illumina and Nanopore library preparation. All 8 subpools were fragmented, size-selected using AMPure XP beads, and Illumina adapters were ligated. The cDNA fragments were cleaned again using beads and amplified, adding the fourth barcode and P5/P7 adapters, followed by size selection and quality check with a Bioanalyzer. Libraries were sequenced with two runs of the Illumina NextSeq 2000 sequencer with P3 200 cycles kits (1.1 billion reads) and paired-end run configuration 140/86/6/0. Libraries with 5% PhiX spike-in were loaded at 1000 pM for one run and 1100 pM for the second run and sequenced to an average depth of # million reads per library.

LR-Split-seq experimental protocol

Nuclei were barcoded and cDNA was purified as specified in the previous section. LR-Split-seq libraries were generated using an input of 200 fmol from the amplified, barcoded Split-seq cDNA before fragmentation (section 2 of the Split-seq protocol). Libraries were built using Oxford Nanopore Technologies Ligation Sequencing Kit (SQK-LSK114) and NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing (E7180L). The Short Fragment Buffer (SFB) from the Ligation Sequencing Kit (SQK-LSK114) during the second wash step. Libraries were loaded on R10.4.1 flowcells (FLO-PRO114M, FLO-MIN114) with an input of 20 fmol and 12 fmol, respectively. Sequencing was done on the GridION and PromethION 2 Solo instruments using the MinKNOW software.

Short-read Split-seq data processing

Fastqs between the two NextSeq runs were concatenated and aligned using kallisto v0.49.0, bustools v0.42.0, and kb python v0.27.3 with the mm39, GENCODE vM32 genome by running the following command:

```
kb count --h5ad --gene-names --sum=nucleus --strand=forward
-r ref/r1_RT_replace.txt -w ref/r1r2r3.txt --kallisto=$kallisto --bustools= $bustools
--workflow= nac -g ref/c57bl6j.t2g -x SPLIT-SEQ -i ref/c57bl6j.idx -t 24 -o kallisto_cou
-c1 ref/c57bl6j.c1 -c2 ref/c57bl6j.c2 next1_B01_13A_R1.fastq.gz next1_B01_13A_R2.fastq.gz.
```

Reference file r1r2r3.txt contains expected barcodes for the 3 rounds and both sets of barcodes, and reference file and r_RT_replace.txt contains oligo dT and corresponding random hexamer barcodes for round 1 barcoding. The resulting counts matrices were merged across the 8 subpools after adding a unique subpool identifier to the 24-nucleotide cell barcodes, then filtered in Seurat v4.1.1 using the following QC metrics across all nuclei: $>$ 500 UMIs,

> 300 genes expressed, < 0.2 doublet score calculated via Scrublet v0.2.3, and $< 0.5\%$ mitochondrial gene expression per nucleus. There were 87,913 final filtered nuclei recovered across all 8 subpools, and 11,013 nuclei for exome capture subpool "13G". For annotation purposes, the exome capture subpool was integrated with the other 7 non-exome capture subpools using Harmony v1.0. Clustered nuclei across all 8 subpools were manually annotated using expression of known marker genes and Seurat label transfer from external datasets for guidance.

LR-Split-seq data preprocessing

The following processing details apply to both the exome capture and non-exome capture data, as well as the MinION and PromethION data.

Reads were basecalled with Nanopore's basecalling software Guppy in super-accurate mode. Adapters were trimmed from reads with Porechop with added custom adapter sequence reflecting the Split-seq adapters in the libraries.

Reads were demultiplexed using LR-Split-pipe[95] with the following settings: `-k WT -c v2 -l1_mm 4 -l2_mm 4 --max_read_len 10000 --max_linker_dist 200`.

Demultiplexed reads were mapped to the mm10 reference genome with minimap2[33] using the following settings: `-ax splice -k14`.

Once again using LR-Split-pipe[95], aligned reads were tagged and filtered for cells with at least 100 identified UMI using the following settings: `-k WT -c v2 --merge_primers --min_umi 100`.

Reads were then run through TranscriptClean[38] with the following settings: `--correctIndels True --correctMismatches False --canonOnly --primaryOnly`.

We used TALON's[36] read labeler with the mm10 reference genome to obtain the genomic sequence around each read's 3' end to flag potential internal priming artifacts. We ran it with the following settings: `--ar 20`.

To call observed transcript isoforms in each cell, reads were annotated with TALON using the union of transcript isoforms reported by the ENCODE4 mouse LR-RNA-seq project and those from GENCODE vM25 for GENCODE vM25 annotated genes[21,101]. TALON was run to use cell barcode tags present in the input alignment files with the following settings: `--cb --l 0 --5p 500 --3p 300 -c 0.8`. Using the ENCODE4 annotation allows us to track transcript isoforms by their TSS, exon junction chain, and TES used. We produced gene and transcript-level AnnData objects without filtering using the `talon_create_adata` utility.

LR-Split-seq exome capture versus non-exome capture comparisons

Using the MinION data, we obtained the read lengths from the read annotation file output from TALON. We obtained the number of exons per transcript from the information TALON provides in the output transcript-level AnnData object. For the transcript novelty assignments, we used the novelty categories for transcripts from our reference as defined by Cerberus from the bulk ENCODE4 mouse LR-RNA-seq dataset[101]. For transcripts that were novel to this reference, we used the TALON novelty categories they were assigned.

We merged cell metadata from the short-read Split-seq data in with the LR-Split-seq data using corresponding cell barcodes. Using Scanpy[102], we filtered out nuclei with < 500 UMI, novel genes, and genes expressed in < 5 nuclei. We normalized the data to 10,000 counts, logged it, and scaled it. We called the top 4,000 highly-variable genes, performed PCA, and found nearest-neighbors using `n_neighbors=10, n_pcs=30, metric=cosine`. We computed the UMAP using default Scanpy parameters and performed Leiden clustering with `resolution=0.25`.

Using PyDESeq2[155], we called differentially-expressed genes between the exome capture and non-exome capture subpools from the unfiltered AnnData directly from TALON (less novel genes), using the different mouse individuals as replicates. We ran the test using default settings. We called significantly differentially-expressed genes with a log2(fold change) threshold of 1 and an adjusted p-value threshold of 0.01. Using our input reference GTF, we determined the total intron length for each transcript. For each gene, we took the median total intron length across all transcripts. We scipy[131] two-sided KS test to determine whether the median total intron length between genes upregulated in the exome capture versus non-exome capture were significantly different.

LR-Split-seq gene-level single-nucleus processing

Using the PromethION data from the exome capture subpool, we merged cell metadata from the short-read Split-seq data in with the LR-Split-seq data using corresponding cell barcodes. Using Scanpy[102], we filtered out nuclei with < 500 UMI, novel genes, and genes expressed in < 5 nuclei. We additionally only retained nuclei that also passed QC in the short-read Split-seq data. We normalized the data to 10,000 counts, logged it, and scaled it. We called the top 4,000 highly-variable genes, performed PCA, and found nearest-neighbors using `n_neighbors=10, n_pcs=30, metric=cosine`. We computed the UMAP using default Scanpy parameters.

We called marker genes from each cell type using Scanpy's `rank_genes_groups` on the log-normalized data. We plotted the top three gene marker results called for each cell type.

LR-Split-seq transcript-level single-nucleus processing

Using the PromethION data from the exome capture subpool, we merged cell metadata from the short-read Split-seq data in with the LR-Split-seq data using corresponding cell barcodes. We filtered out all transcripts that were novel with respect to the ENCODE4 bulk mouse LR-RNA-seq dataset and transcripts belonging to novel genes.

Given the low remaining read depth per nucleus after filtering out all novel transcripts unique to the LR-Split-seq data, we used Scanpy[102] to filter out nuclei with < 60 UMI. We further filtered out transcripts expressed in just one nucleus. We normalized the data to 10,000 counts, logged it, and scaled it. We called the top 10,000 highly-variable transcripts, performed PCA, and found nearest-neighbors using `n_neighbors=10, n_pcs=30, metric=cosine`.

We called marker transcripts from each cell type using Scanpy's `rank_genes_groups` on the log-normalized data. We plotted the top three transcript marker results called for each cell type.

LR-Split-seq marker transcript analysis

Using the markers reported from the previous step, we called the highest marker transcript for each cell type / gene combination using the `scores` column in the `rank_genes_groups` output. That is to say, for every gene / cell type combination we retained the single best-scoring marker transcript. We then counted up how many unique best-scoring marker transcripts there were across the cell types, and determined how many genes had more than one best-scoring marker transcript across cell types.

LR-Split-seq DU tests

For all of the following, we performed all DU tests (DU-T, DU-TSS, DU-EC, and DU-TES). All DU tests were done using Swan's `die_gene_test` function[56]. Briefly, this test compares the pseudobulked percent isoform (π) values for each transcript in a gene between two conditions. It creates an $nx2$ table (where $n =$ the number of transcripts) and performs a χ^2 test on this table, which results in the p-value for the switch for the gene. To quantify effect size, DPI ($\Delta\pi$) values are calculated as the sum of the difference between π values for each transcript for the top two highest π values with the same sign. Significant switches were called with a $\Delta\pi$ threshold of ≥ 10 and an adjusted p-value threshold of ≤ 0.05. For each test, we computed the fraction of DU events that were called as DU-T, DU-TSS, DU-EC, or DU-TES.

Inter-cell type tests

DU tests were performed between each unique pair of cell types.

Inter sub-cell type tests

We considered only our sub cell types for those that had at least 50,000 reads total across the constituent population of nuclei. We performed DU tests between each unique pair of sub cell types.

Intra-cell type genotype tests

We performed DU tests between C57BL/6J and CAST/EiJ within each cell type in our LR-Split-seq dataset.

Inter-cell type genotype tests

We performed DU tests between C57BL/6J and CAST/EiJ globally across our LR-Split-seq dataset.

Intra-cell type sex tests

We performed DU tests between male and female mice within each cell type in our LR-Split-seq dataset.

Inter-cell type sex tests

We performed DU tests between male and female mice globally across our LR-Split-seq dataset.

ENCODE4 bulk LR-RNA-seq inter-tissue DU tests

We obtained expression and percent isoform usage data from the ENCODE4 mouse LR-RNA-seq dataset. We limited the datasets used to just the C57BL/6J / CAST/EiJ F1 hybrid mice at 2 months of age, which is the closest represented time point to the mice from our LR-Split-seq dataset. Using the same strategy as described in the previous section, we performed DU tests between each unique pair of the five tissues: hippocampus, cortex, adrenal gland, skeletal muscle, and heart.

Bulk and single-nucleus cell type-specific transcript diversity

For our LR-Split-seq data, we used Cerberus'[101] `get_subset_triplets` to compute the transcript diversity across genes for each transcript expressed in each cell type separately. Additionally, we used the `sample_det` triplets from the ENCODE4 human bulk LR-RNA-seq dataset for the PGP1-derived excitatory neurons and astrocytes to compare diversity in our LR-Split-seq cell types directly to analogous bulk data.

Characterization of genes with high transcript diversity

We obtained a list the top 100 genes per cell type by number of transcript isoforms expressed. We performed gene ontology (GO) term enrichment analysis using the R package EnrichR[133] to query the "GO_Biological_Process_2021" database.

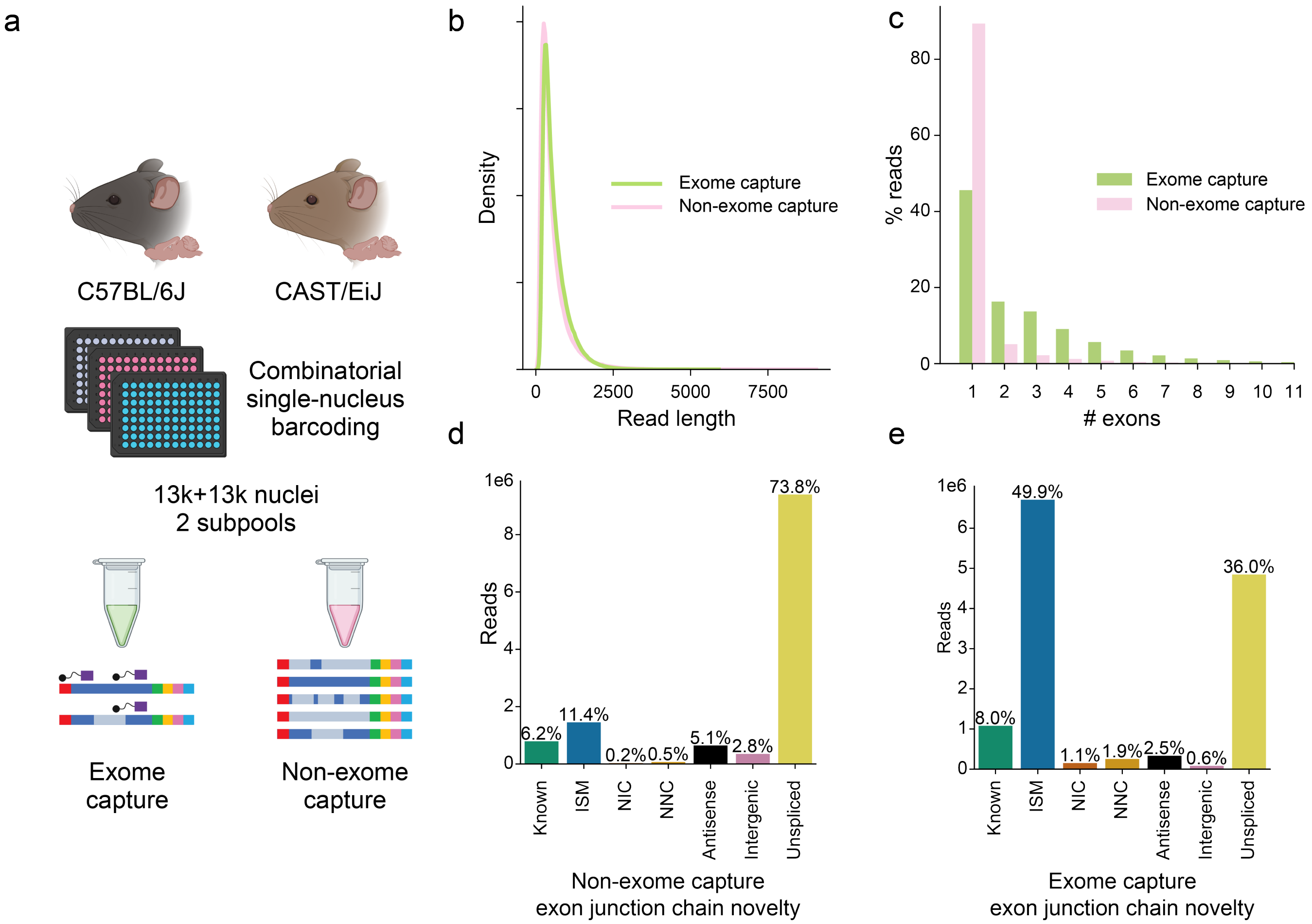


Figure 4.1: **Exome capture in single nuclei improves spliced RNA yield. a,** Experimental overview for comparison of exome capture vs. non exome capture LR-Split-seq libraries. **b,** Kernel density estimations for read length distributions by capture strategy. **c,** Percentage of demultiplexed reads by number of exons in each read between exome and non-exome capture. **d-e,** Percentage of demultiplexed reads by novelty category in **d,** non-exome capture and **e,** exome capture.

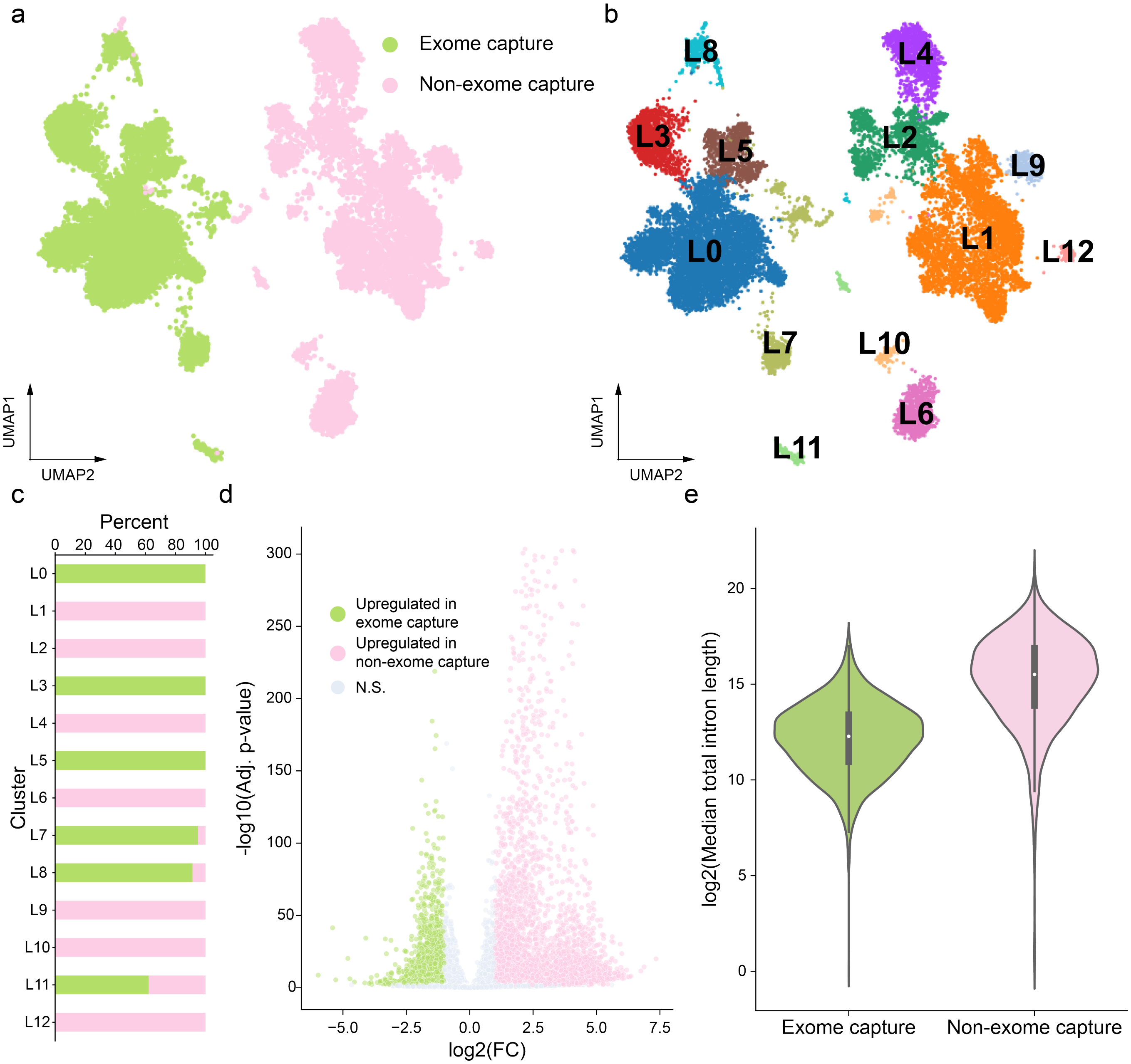


Figure 4.2: **Upregulated genes in non-exome capture are enriched for longer total intron length. a-b,** UMAP of MinION LR-Split-seq data colored by **a,** capture strategy and **b,** unbiased cluster assignments. **c,** Percentage of nuclei assigned to each cluster by capture strategy. **d,** Volcano plot of differentially expressed genes when comparing exome capture to non-exome capture. **e,** Median total intron length for genes upregulated in exome capture vs. non-exome capture.

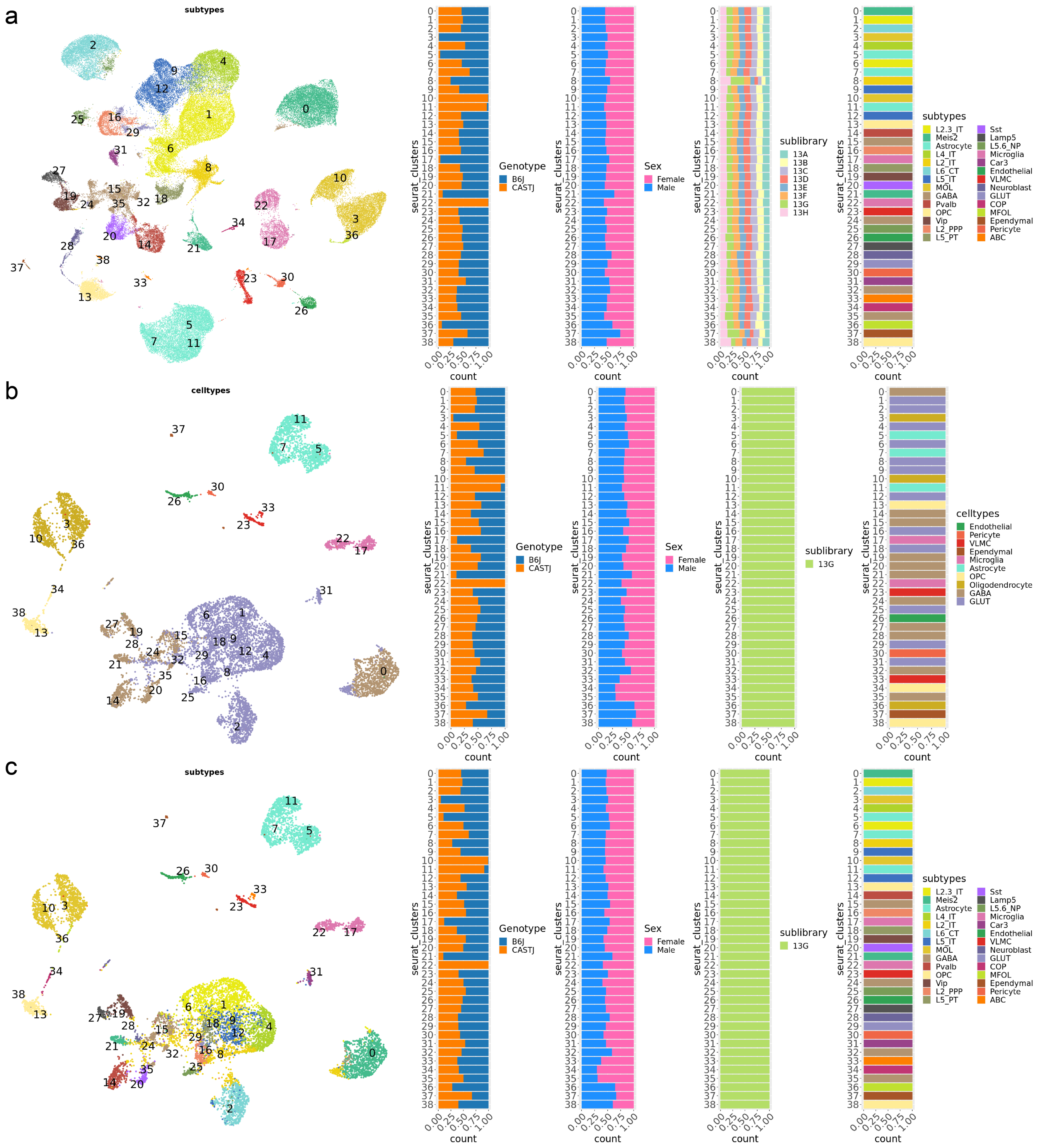


Figure 4.3: **Short-read Split-seq recovers canonical cortical cell types and subtypes. a,** UMAP of short-read Split-seq unbiased clustering on nuclei and proportions of nuclei per genotype, sex, subpool, and sub-cell type per cluster. **b,** UMAP of short-read Split-seq general cell type assignments for exome capture supbool only. Proportions of nuclei per genotype, sex, and cell type per cluster. **c,** UMAP of short-read Split-seq sub cell type assignments for exome capture supbool only. Proportions of nuclei per genotype, sex, and sub cell type per cluster.

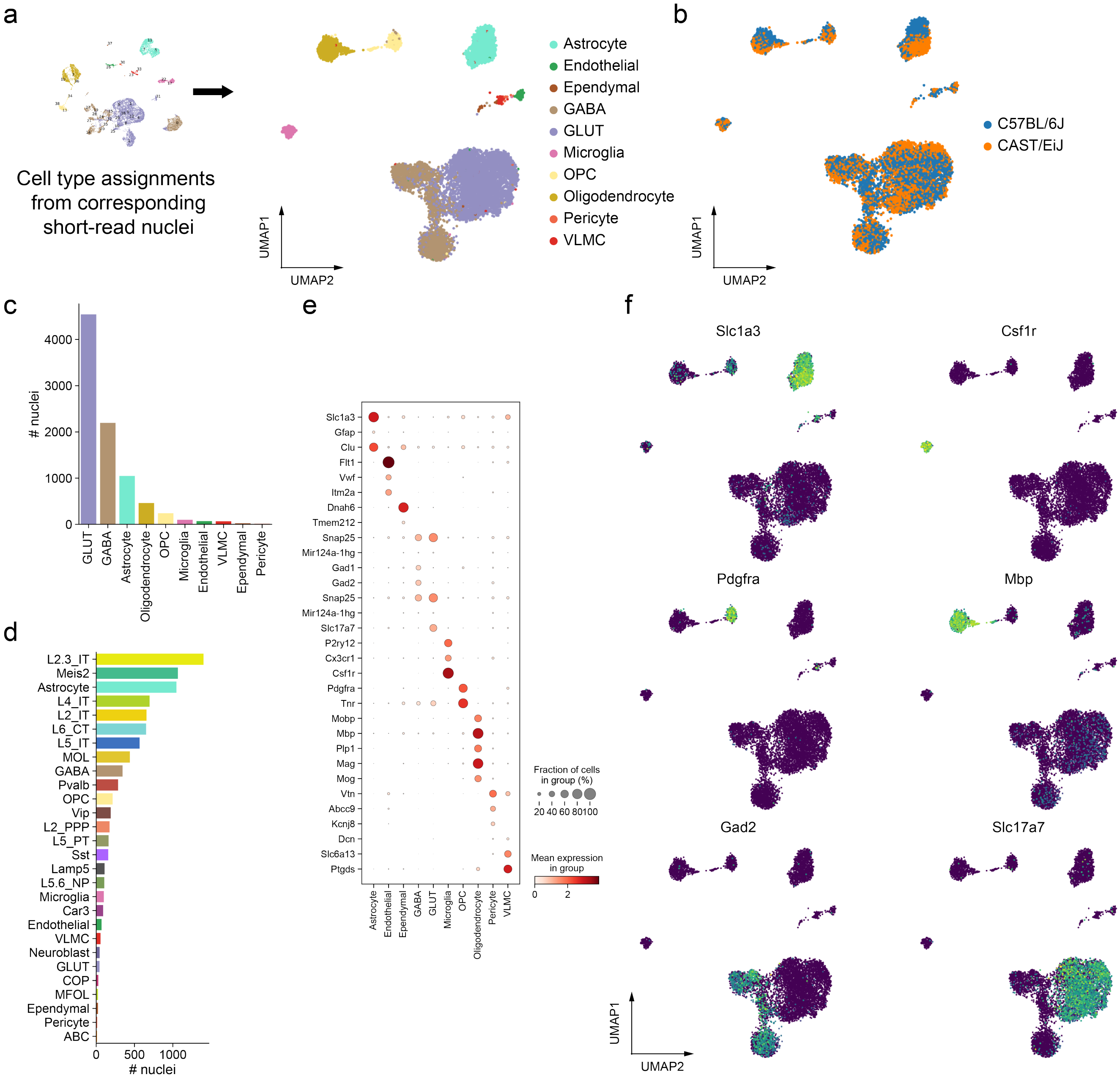
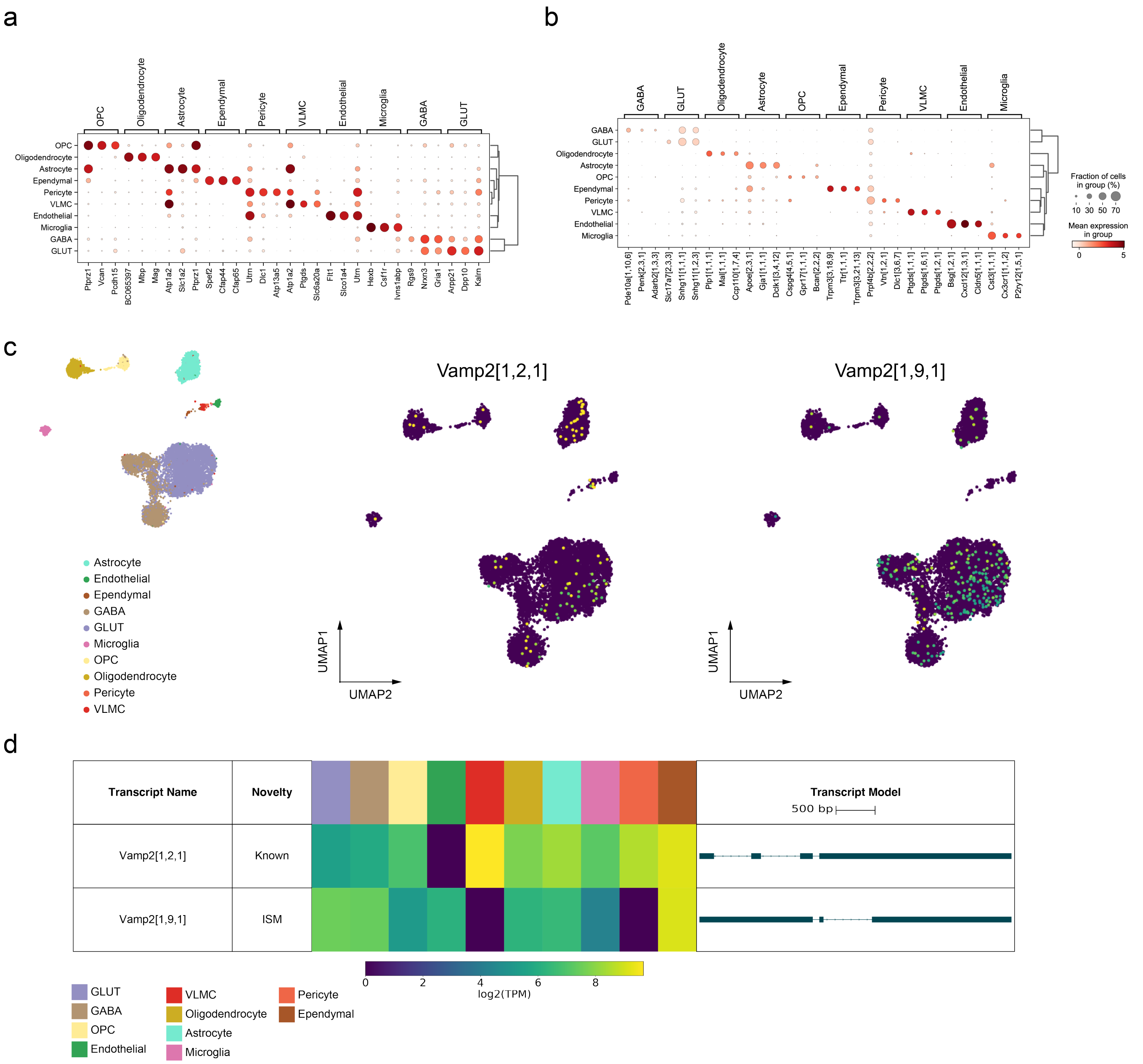


Figure 4.4: **LR-Split-seq captures distinct cell types in mouse cortex. a,** UMAP of gene-level exome capture PromethION LR-Split-seq colored by cell types as annotated in the corresponding short-read Split-seq cells. **b,** UMAP of LR-Split-seq data colored by genotype. **c,** Number of nuclei recovered per cell type from the transcript-level LR-Split-seq data. **d,** Number of nuclei recovered per sub cell type from the transcript-level LR-Split-seq data. **e,** Expression of marker genes for cortical cell types from literature. **f,** Marker gene expression for different cell types in the gene-level LR-Split-seq UMAP.



Transcript Name	Novelty	Transcript Model
Vamp2[1,2,1]	Known	
Vamp2[1,9,1]	ISM	

Figure 4.5: **Unbiased marker gene and transcript analysis from LR-Split-seq. a,** Top 3 marker genes from each cell type in gene-level LR-Split-seq. **b,** Top 3 marker transcripts from each cell type in transcript-level LR-Split-seq. **c,** Expression of two recovered marker transcripts of *Vamp2* plotted on the gene-level UMAP. These transcripts were called as markers of astrocytes and glutamatergic neurons respectively. **d,** Expression and transcript models of marker *Vamp2* transcripts in each cell type.

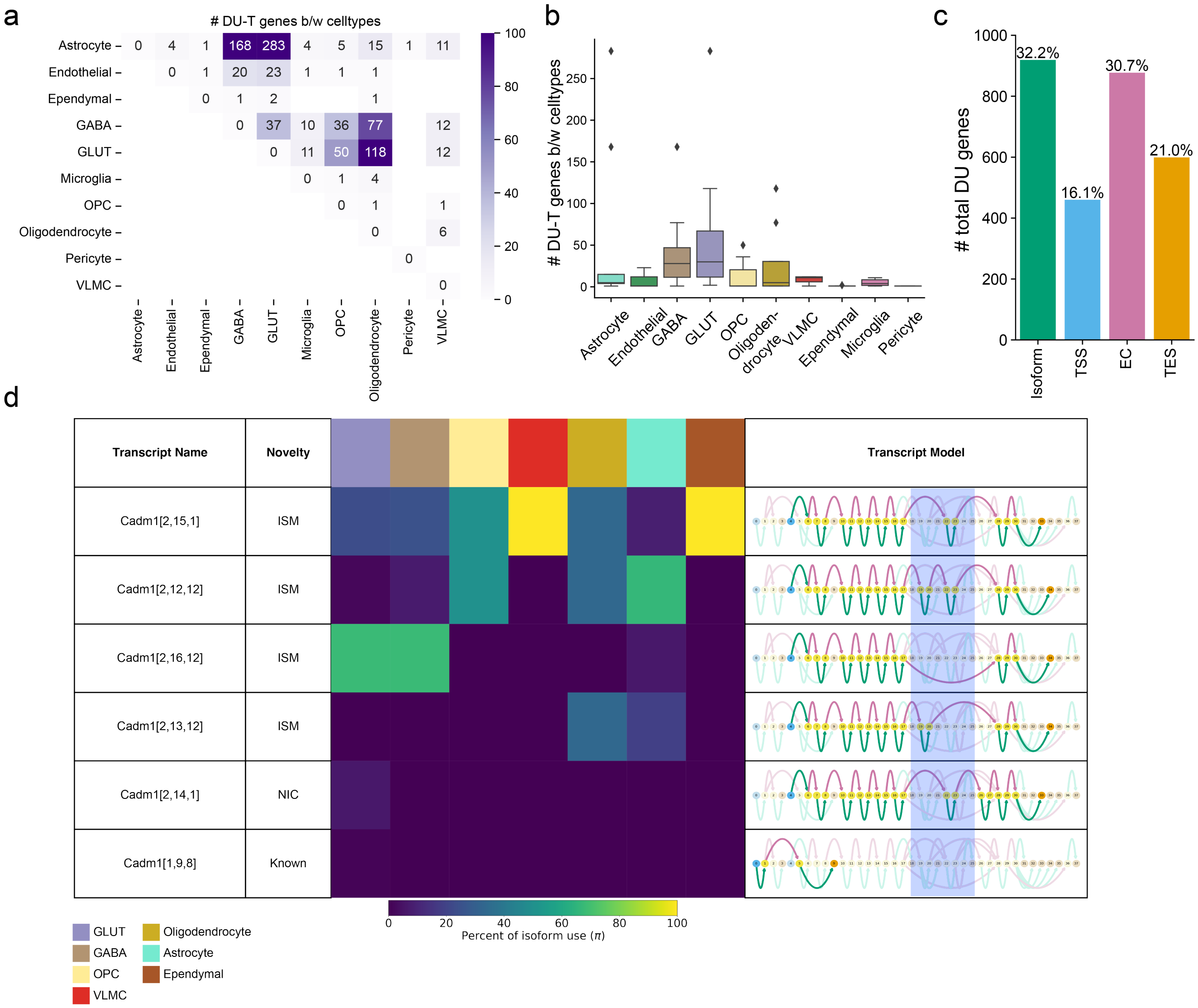


Figure 4.6: **Pairwise cell type DU-T analysis reveals cell type-specific transcript isoform usage. a,** Number of DU-T genes between each pair of cell types. **b,** Number of DU-T genes between each cell type and every other cell type. **c,** Total number of DU genes called on the whole transcript, TSS, EC, and TES level across all comparisons. **d,** Swan report of DU gene *Cadm1*, which displays exon skipping (highlighted) in the neuronal cell types compared to the astrocytes.

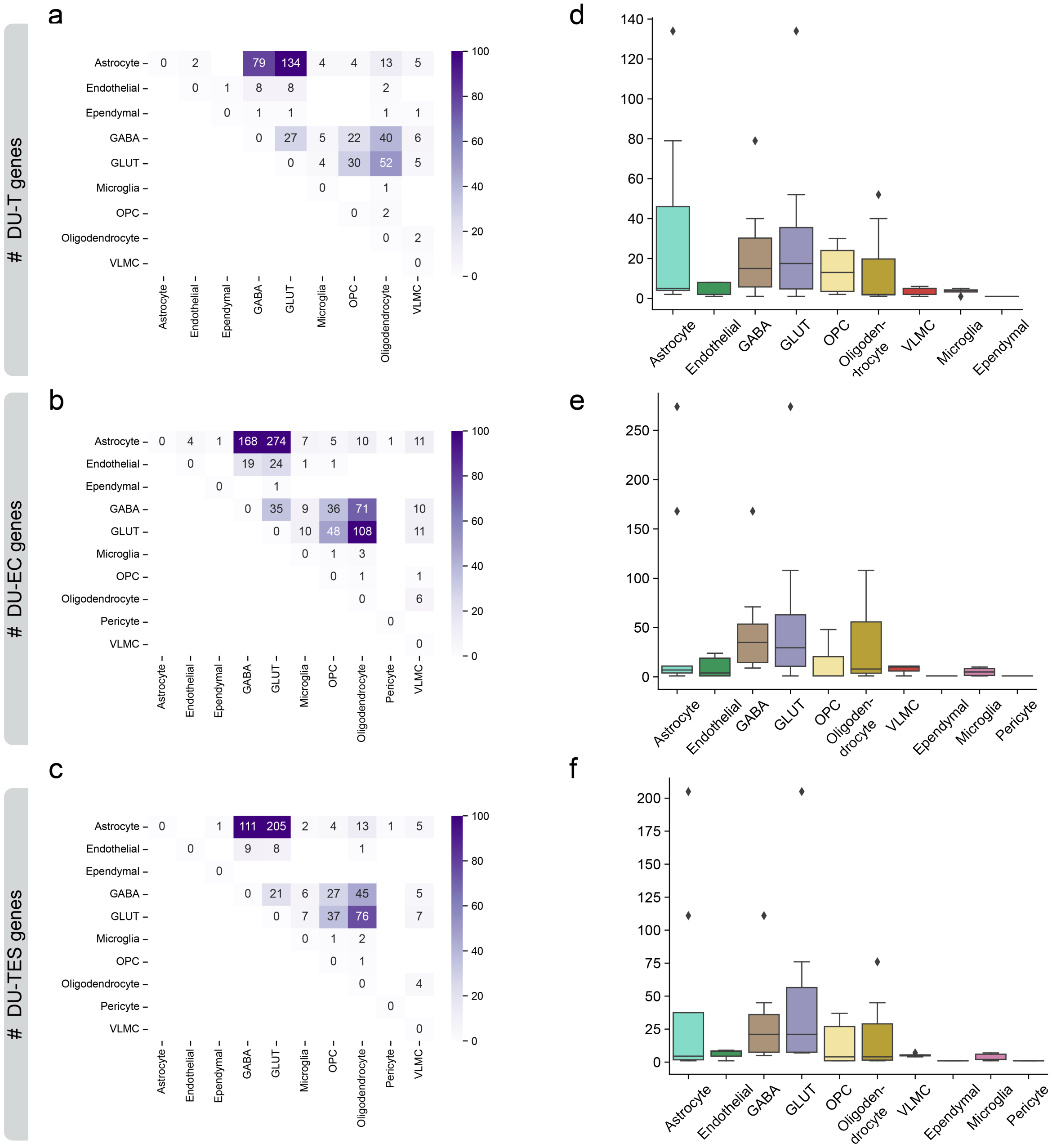


Figure 4.7: **Pairwise cell type DU-TSS, DU-EC, and DU-TES analysis reveals cell type-specific usage. a-c,** Number of **a,** DU-TSS **b,** DU-EC **d,** DU-TES usage genes between each pair of cell types. **d-f,** Number of **d,** DU-TSS **b,** DU-EC **d,** DU-TES usage genes between each cell type and every other cell type.

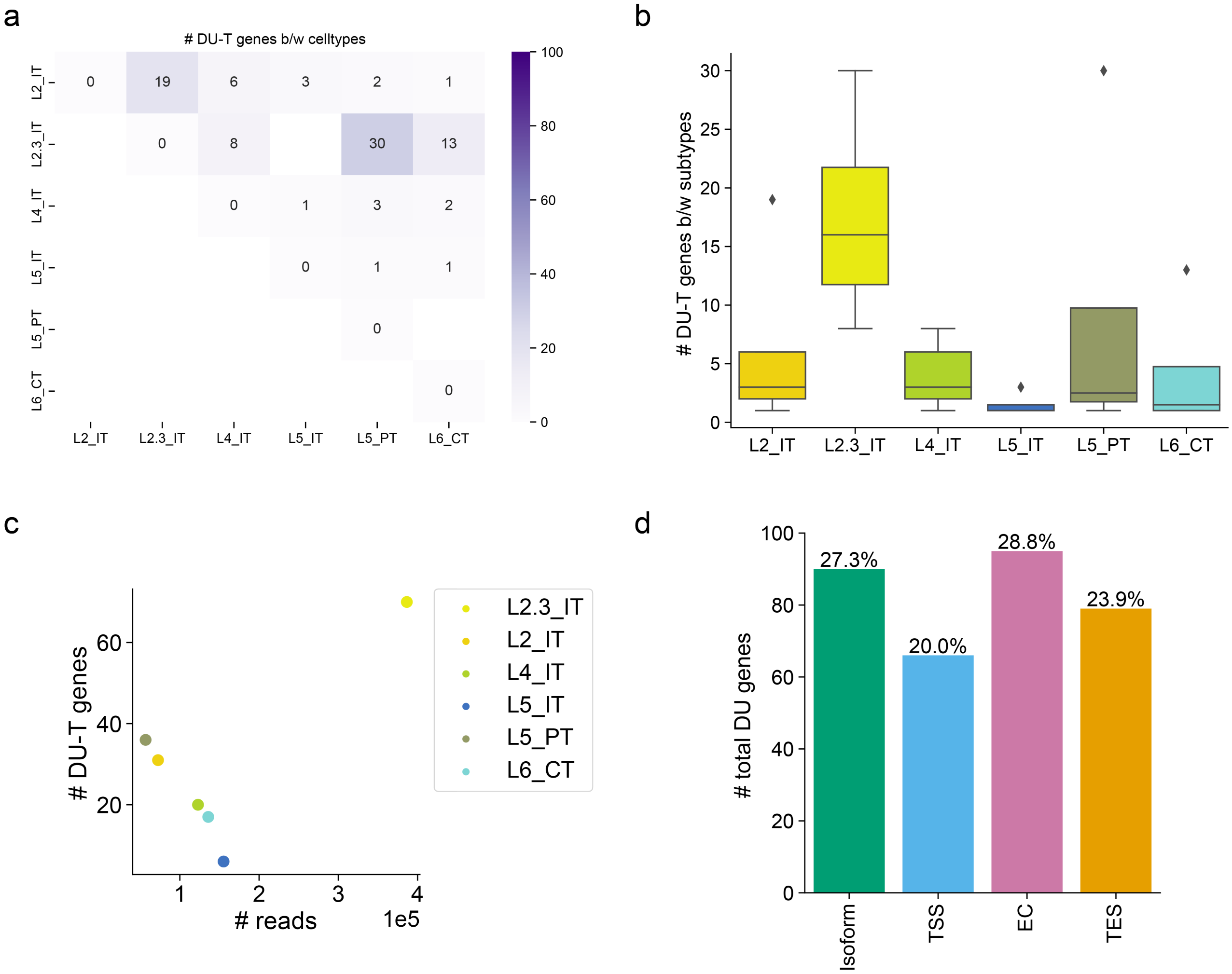


Figure 4.8: **Pairwise excitatory neuron subtype DU-T analysis reveals neuronal subtype-specific transcript isoform usage. a,** Number of DU-T genes between each pair of excitatory neuron subtypes. **b,** Number of DU-T genes between each excitatory neuron subtype and every other excitatory neuron subtype. **c,** Total number of reads per excitatory neuron subtype versus the total number DU-T genes. **d,** Total number of DU genes called on the whole transcript, TSS, EC, and TES level across all comparisons.

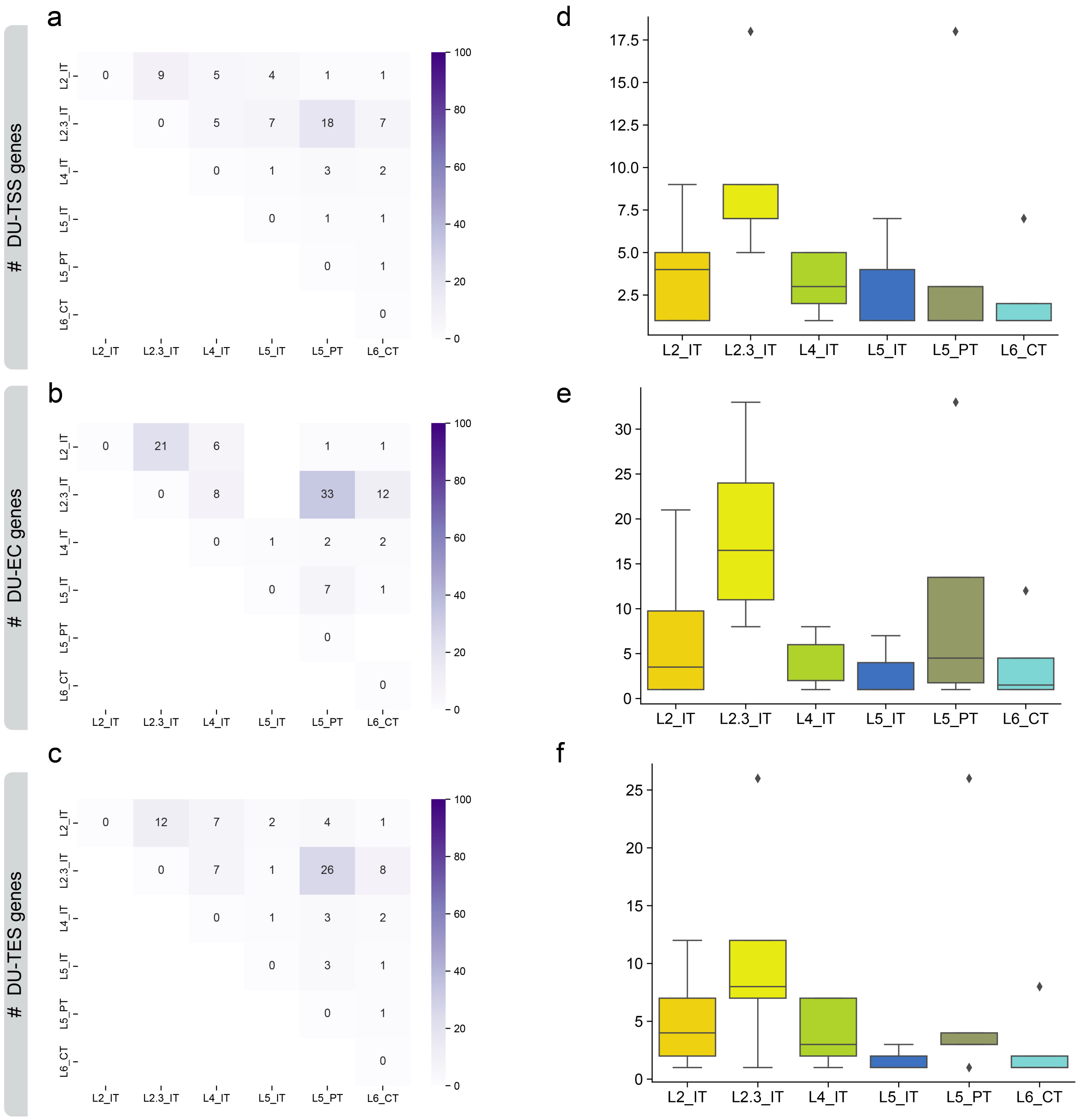


Figure 4.9: **Pairwise cell type DU-TSS, DU-EC, and DU-TES analysis reveals excitatory neuron subtype-specific usage. a-c,** Number of **a,** DU-TSS **b,** DU-EC **d,** DU-TES usage genes between each pair of excitatory neuron subtypes. **d-f,** Number of **d,** DU-TSS **b,** DU-EC **d,** DU-TES usage genes between each excitatory neuron subtype and every other excitatory neuron subtype.

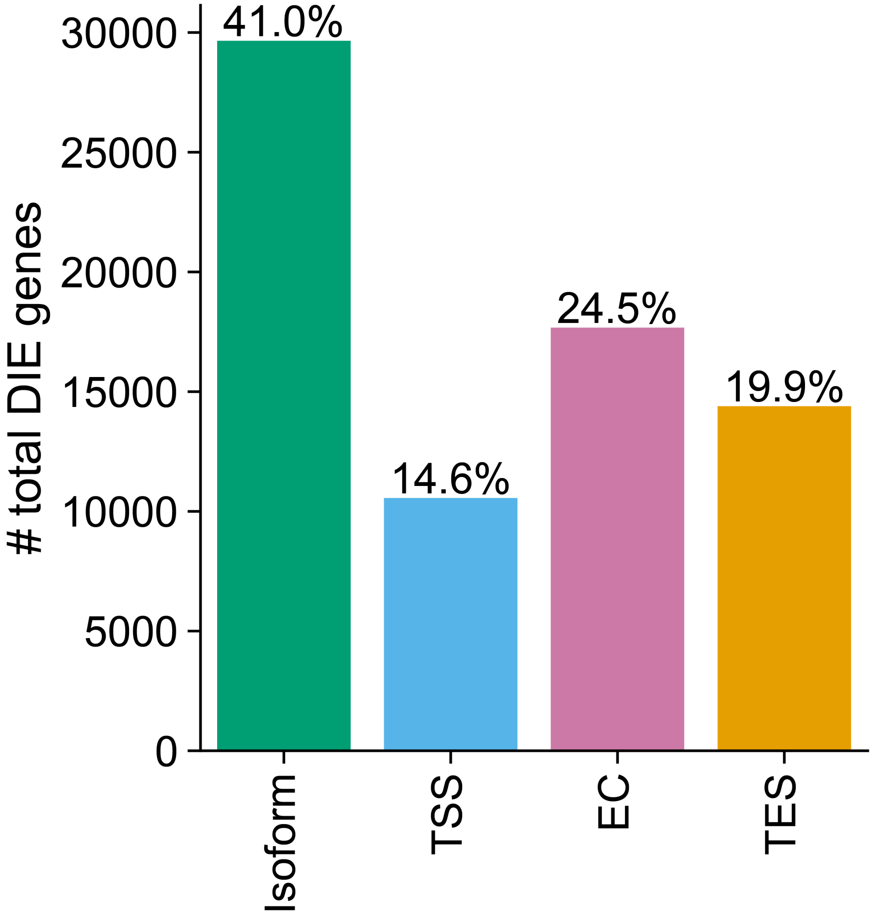


Figure 4.10: **Number of DU events between tissues in the bulk ENCODE4 LR-RNA-seq mouse dataset.**

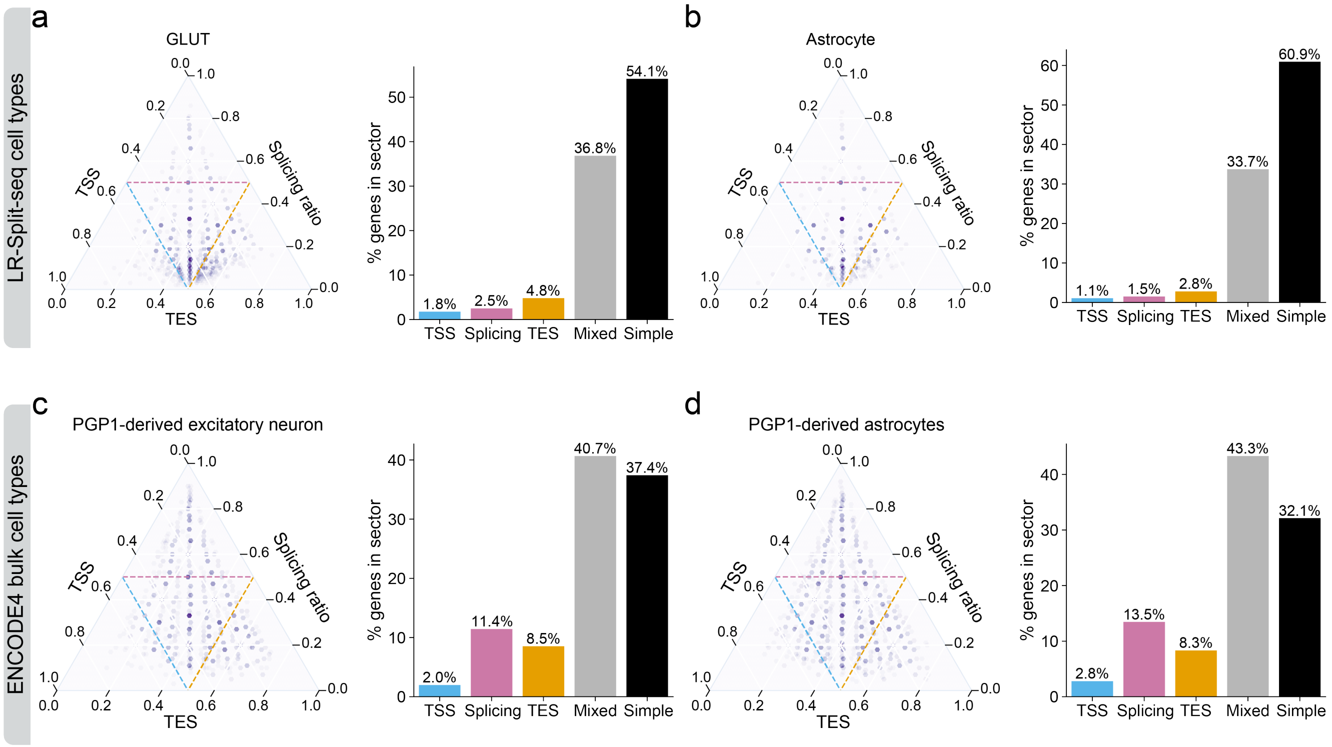


Figure 4.11: **Transcript diversity from cell types in LR-Split-seq does not reflect transcript diversity from bulk cell line data. a-b,** Gene structure simplices and proportion of genes per sector assignments as described in Reese et al.[101] for LR-Split-seq transcripts observed in **a,** glutamatergic neurons and **b,** astrocytes. **c-d** Gene structure simplices and proportion of genes per sector assignments from the ENCODE4 bulk human LR-RNA-seq dataset for transcripts observed in **c,** PGP1-derived excitatory neurons and **d,** PGP1-derived astrocytes.

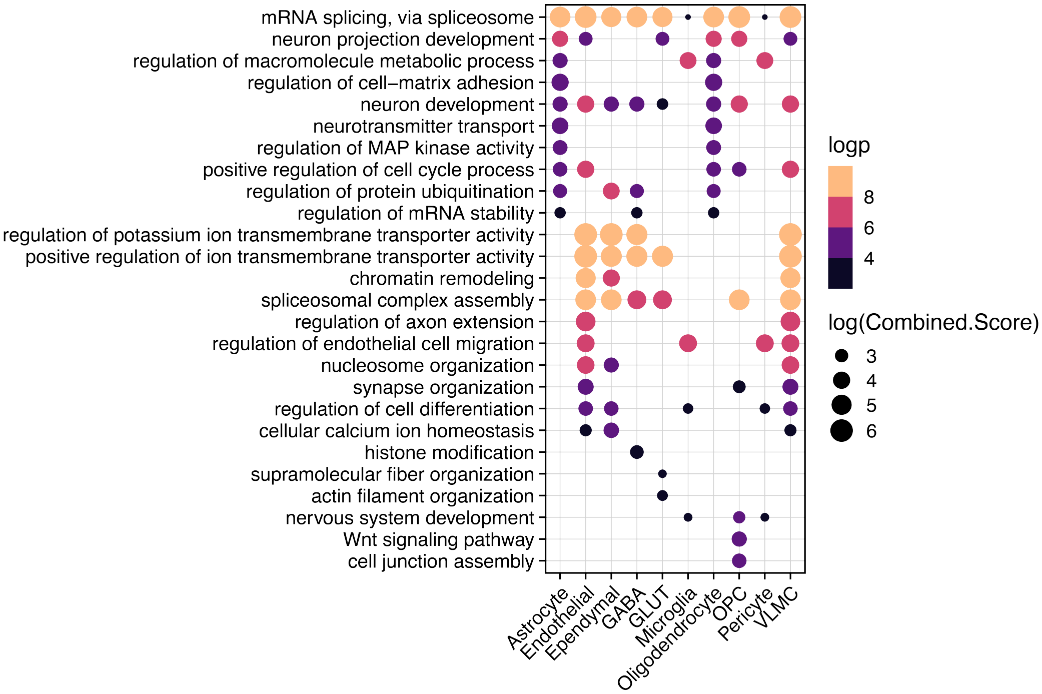
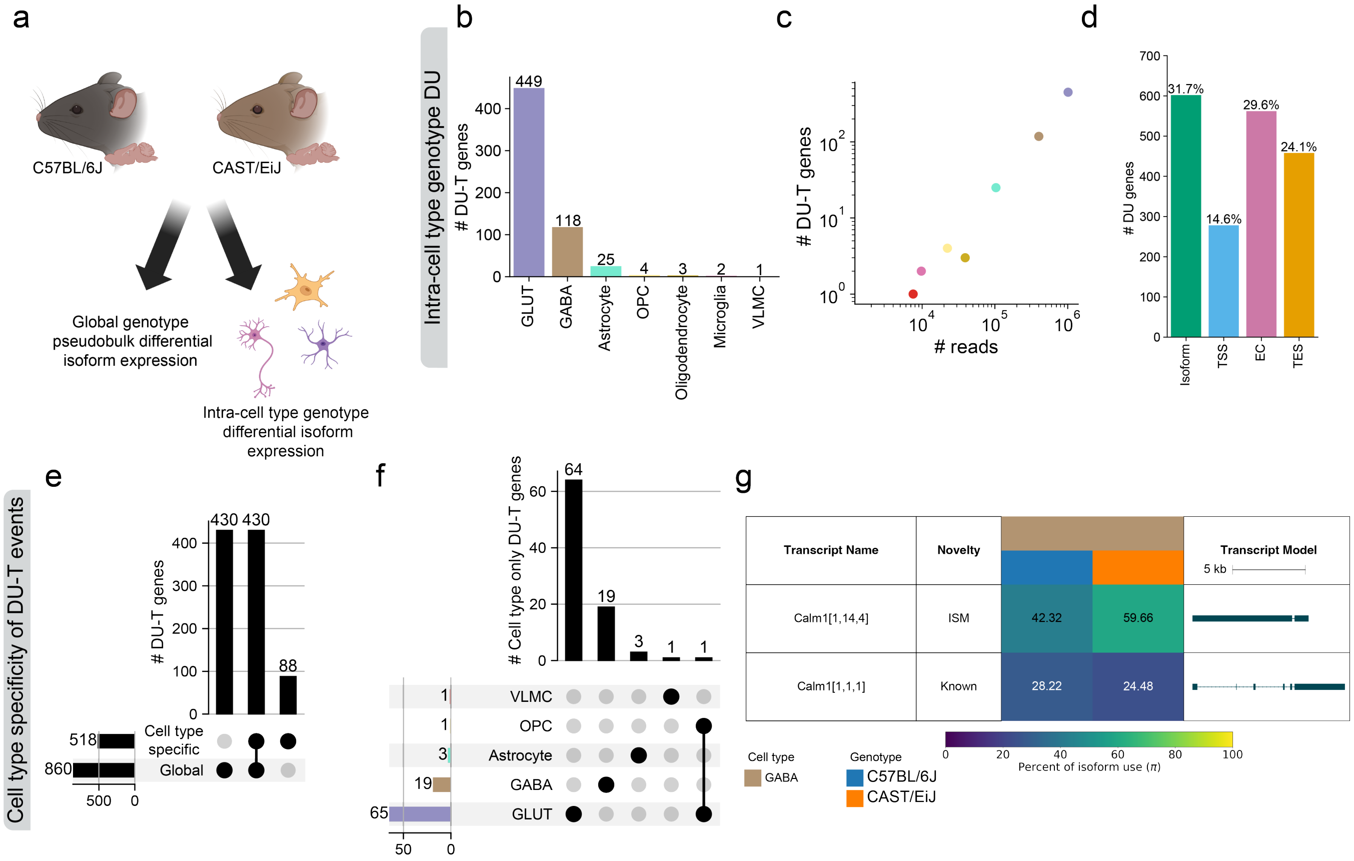


Figure 4.12: **Genes with high transcriptional diversity in distinct cell types demonstrate enrichment for neural processes.** Selected GO terms associated with the top 100 transcriptionally-diverse genes in each cell type.



Transcript Name	Novelty	C57BL/6J	CAST/EiJ	Transcript Model
Calm1[1,14,4]	ISM	42.32	59.66	
Calm1[1,1,1]	Known	28.22	24.48	

Figure 4.13: **LR-Split-seq on two distinct mouse strains reveals cell type-specific genotype DU-T events. a,** Overview of comparisons. DU tests were performed between genotypes globally and within each cell type. **b,** Number of DU-T genes called in each cell type between genotypes. **c,** Number of reads per cell type versus the number of called DU-T genes between genotypes. **d,** Total number of DU-T, DU-TSS, DU-EC, and DU-TES genes called between genotypes in each cell type. **e,** Intersections of DU-T genes called globally and in specific cell types. **f,** Overlap of cell type only DU-T genes by cell type they were called in. **g,** Swan report of DU-T gene *Calm1*, called in the cell type comparison but not globally, in the GABAergic neurons.

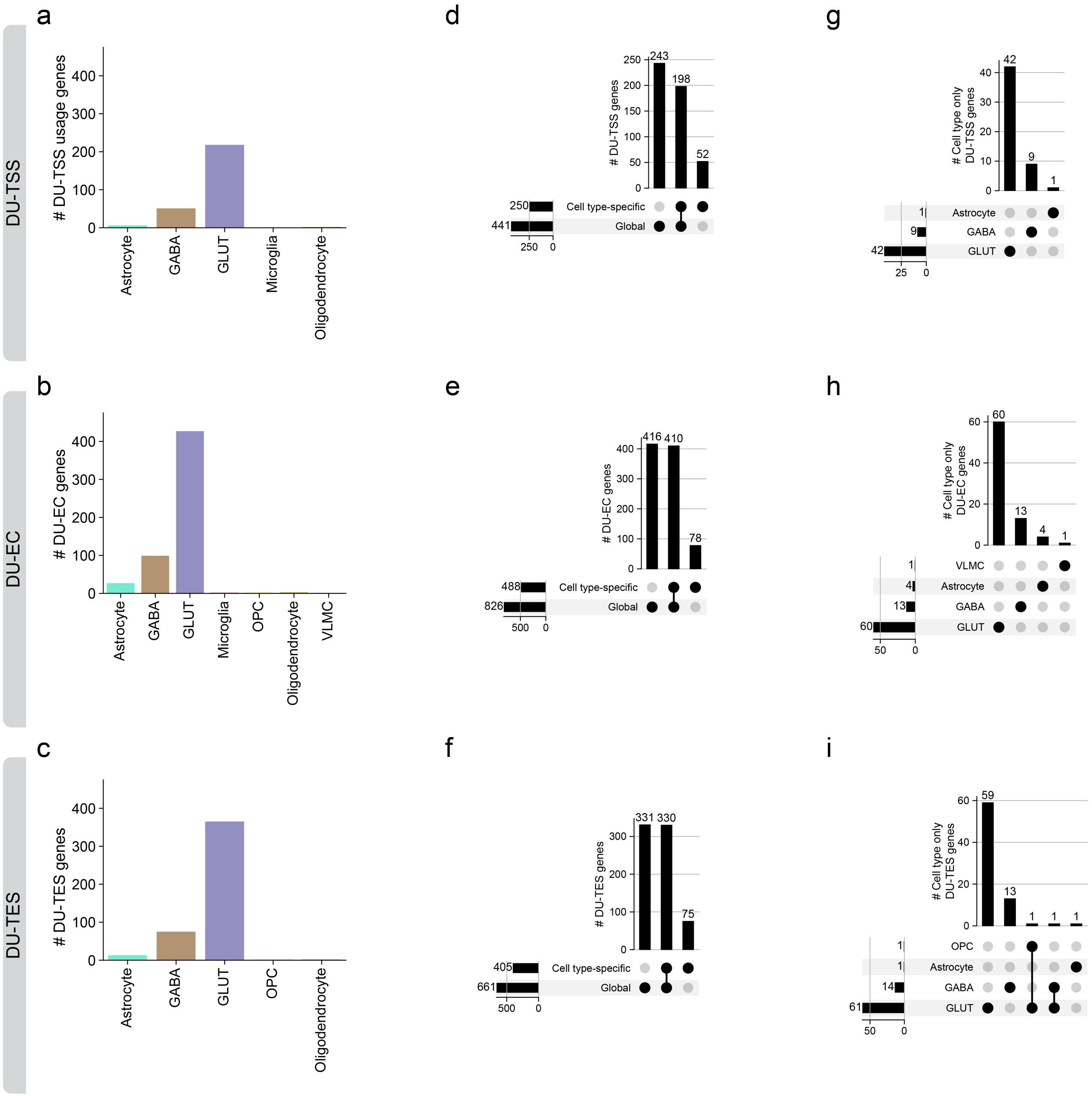
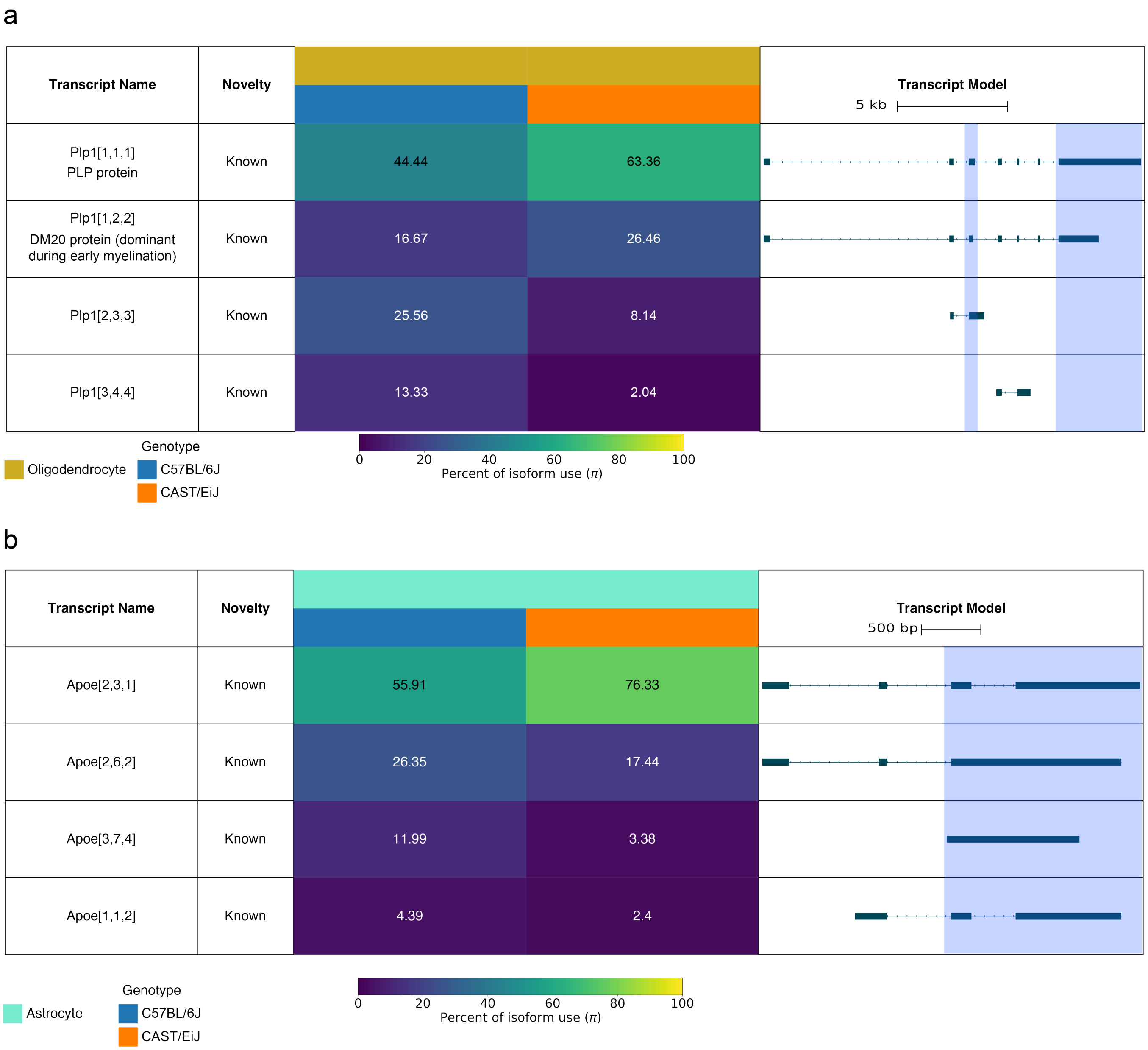


Figure 4.14: **LR-Split-seq on two distinct mouse strains reveals cell type-specific DU-TSS, DU-EC, and DU-TES events between genotypes. a-c,** Number of **a,** DU-TSS **b,** DU-EC and **c,** DU-TES genes in each cell type between genotypes. **d-f,** Intersections of **d,** DU-TSS **e,** DU-EC and **f,** DU-TES genes called globally and in specific cell types. **g-i,** Overlap of cell type only **g,** DU-TSS **h,** DU-EC and **i,** DU-TES genes by cell type they were called in.



Transcript Name	Novelty	C57BL/6J	CAST/EiJ
Plp1[1,1,1] PLP protein	Known	44.44	63.36
Plp1[1,2,2] DM20 protein (dominant during early myelination)	Known	16.67	26.46
Plp1[2,3,3]	Known	25.56	8.14
Plp1[3,4,4]	Known	13.33	2.04

Transcript Name	Novelty	C57BL/6J	CAST/EiJ
Apoe[2,3,1]	Known	55.91	76.33
Apoe[2,6,2]	Known	26.35	17.44
Apoe[3,7,4]	Known	11.99	3.38
Apoe[1,1,2]	Known	4.39	2.4

Figure 4.15: **DU-T between genotypes in *Plp1* and *Apoe*. a,** Swan report for transcripts of *Plp1*, which was called as DU-T between genotypes globally and specifically in the oligodendrocytes. Differences between transcripts highlighted in blue. **b,** Swan report for transcripts of *Apoe*, which was called as DU-T between genotypes globally and specifically in the astrocytes.

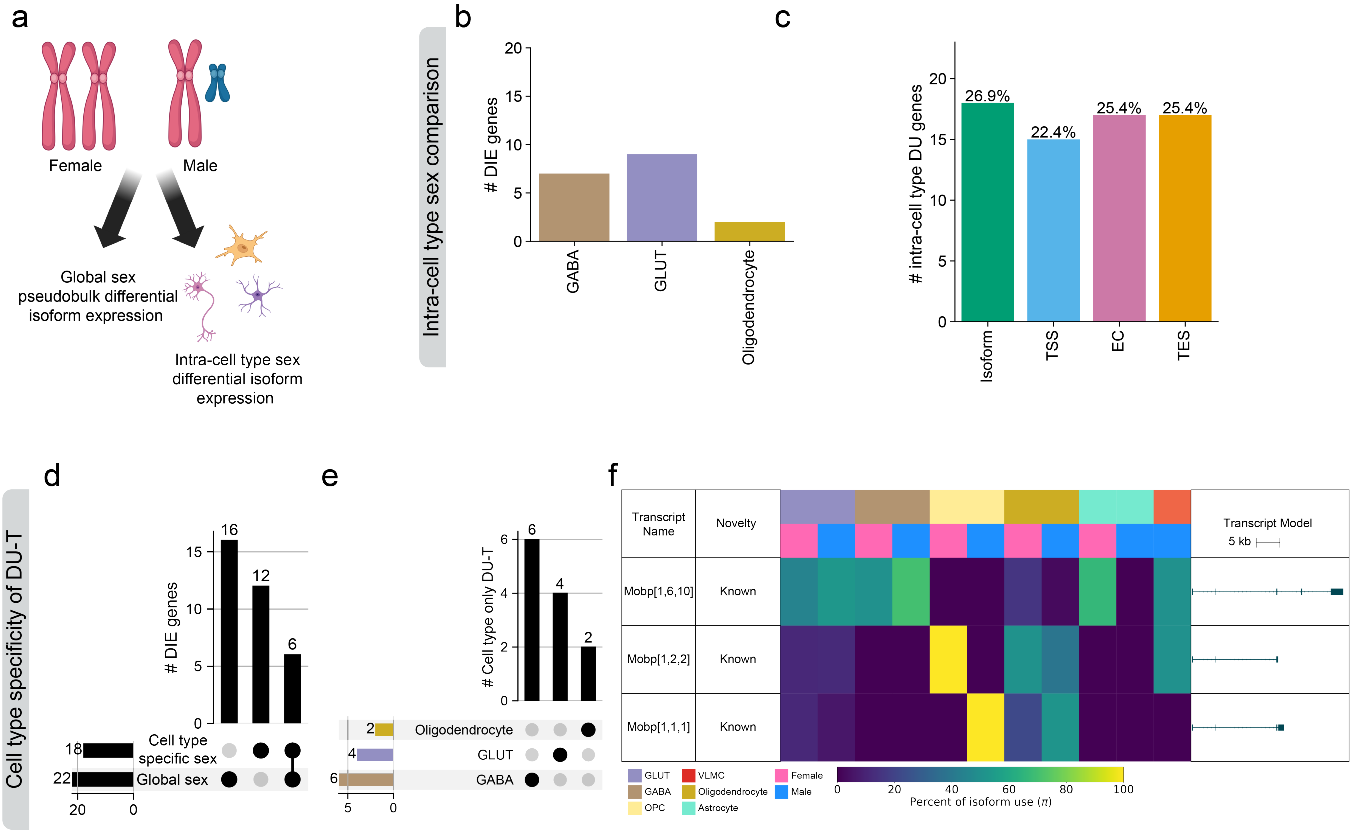


Figure 4.16: **LR-Split-seq on two distinct mouse strains reveals cell type-specific DU-T events between sexes. a,** Overview of comparisons. DU tests were performed between sexes globally and within each cell type. **b,** Number of DU-T genes called in each cell type between sexes. **c,** Total number of DU genes called between sexes in each cell type on the whole transcript, TSS, EC, and TES level. **d,** Intersections of DU-T genes called globally and in specific cell types. **e,** Overlap of cell type only transcript level DU-T genes by cell type they were called in. **f,** Swan report of DU-T gene *Mobp*, called between sexes only in the oligodendrocytes.

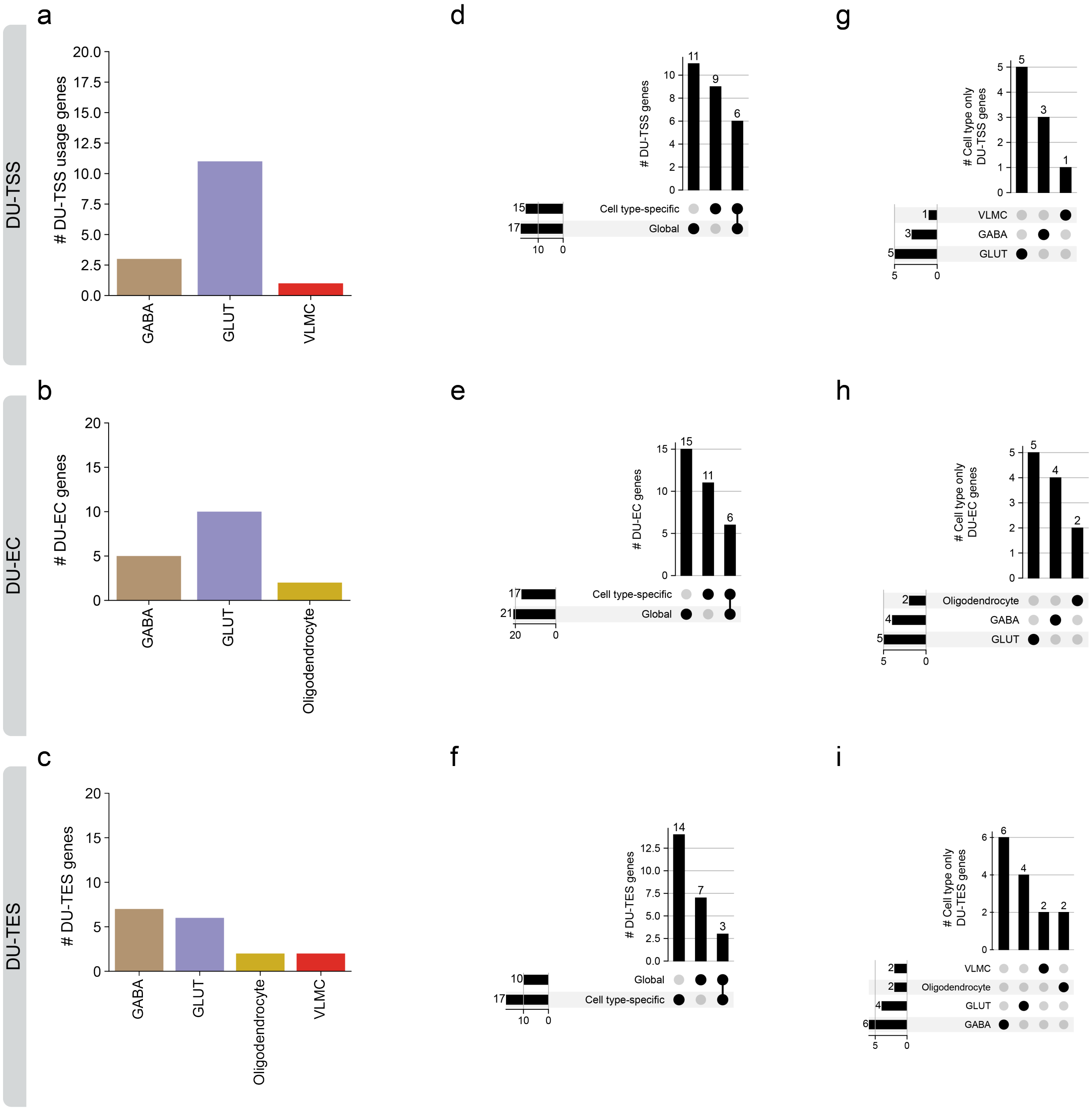


Figure 4.17: **LR-Split-seq reveals cell type-specific DU-TSS, DU-EC, and DU-TES events between sexes. a-c,** Number of **a,** DU-TSS **b,** DU-EC and **c,** DU-TES genes in each cell type between sexes. **d-f,** Intersections of **d,** DU-TSS **e,** DU-EC and **f,** DU-TES genes called globally and in specific cell types. **g-i,** Overlap of cell type only **g,** DU-TSS **h,** DU-EC and **i,** DU-TES genes by cell type they were called in.

Platform	Exome capture	Non-exome capture
Nanopore MinION	X	X
Nanopore PromethION	X	
Illumina NextSeq2000	X	X

Table 4.1: Sequencing platforms used for exome capture and non-exome capture subpools.

Cell type label	Description	Number of nuclei
GLUT	Glutamatergic (excitatory) neuron	4542
GABA	GABAergic (inhibitory) neuron	2195
Astrocyte	Astrocyte	1047
Oligodendrocyte	Oligodendrocyte	462
OPC	Oligodendrocyte precursor cell	241
Microglia	Microglia	99
Endothelial	Endothelial	70
VLMC	Vascular lepotomeningeal cell	66
Ependymal	Ependymal	23
Pericyte	Pericyte	13

Table 4.2: General cell types, their corresponding labels, and number of nuclei called from transcript-level PromethION LR-Split-seq data.

Sub cell type label	Description	Number of nuclei
L2.3_IT	Layer 2/3 intratelencephalic neuron	1400
Meis2	Medium spiny neuron	1066
Astrocyte	Astrocyte	1047
L4_IT	Layer 4 intratelencephalic neuron	696
L2_IT	Layer 2 intratelencephalic neuron	656
L6_CT	Layer 6 corticothalamic neuron	651
L5_IT	Layer 5 intratelencephalic neuron	566
MOL	Mature oligodendrocyte	439
GABA	General GABAergic neuron	344
Pvalb	Pvalb+ inhibitory neuron	285
OPC	Oligodendrocyte precursor	214
Vip	Vip+ inhibitory neuron	189
L2_PPP	Layer 2 post-pre-par subiculum	174
L5_PT	Layer 5 pyramidal tract neuron	161
Sst	Sst+ inhibitory neuron	157
Lamp5	Lamp+ inhibitory neuron	109
L5.6_NP	Layer 5/6 near-projecting neuron	105
Microglia	Microglia	99
Car3	Car3+ excitatory neuron	90
Endothelial	Endothelial	70
VLMC	Vascular lepotomeningeal cell	56
Neuroblast	Neuroblast	45
GLUT	General glutamatergic neuron	43
COP	Committed oligodendrocyte precursor	27
MFOL	Myelin-forming oligodendrocytes.	23
Ependymal	Ependymal	23
Pericyte	Pericyte	13
ABC	Slc47a1+ VLMC	10

Table 4.3: Sub cell types, their corresponding labels, and number of nuclei called from transcript-level PromethION LR-Split-seq data.

Chapter 5

Future directions

Improving variant-aware transcript identification with pantranscriptomes

Though the field of analyzing full-length transcriptomes has dramatically matured through the last few years, there are certainly clear avenues to continue advancing the field. Recent pangenome projects, recognizing the importance of being able to incorporate genetic variation and diversity into any genomics workflow, have incorporated multiple individuals as sources of genetic variability both at the single nucleotide polymorphism (SNP) and structural variant (SV) level[156]. In contrast to linear reference genome based approaches, pangenomes typically represent genomic sequence using graph models where any given walk through the graph represents a valid haplotype[157]. Therefore pangenomes simultaneously represent any possible haplotype from the cohort of individuals that was used to generate it. Representing genetic diversity using pangenomes has already improved variant calling (particularly for SVs) and reference mapping tasks on a variety of genomics data[156]. A clear next step forward is applying a similar philosophy to generate a pantranscriptome, which would improve efforts to call transcripts with sensitivity to the variation in the genome they derive from. This would provide a clear benefit for tasks where genetic variation is important to consider, such as variant effect prediction in patient transcriptomes where pathological variants might be overlooked due to their transcriptomic context in current reference annotations[158]. While pantranscriptomic methods have been developed to generate pantranscriptomes, map short-read RNA-seq data to said pantranscriptomes, and quantify haplotype-specific transcript expression, these tools are new and limited in scope. Notably, an unsolved problem is determining which haplotype-specific transcripts are equivalent to one another. While perhaps simple if the difference in haplotype between transcripts is a simple SNP, this becomes more complicated if the difference contains structural variation[159]. Furthermore, there is less interest in creating pangenomes for model organisms such as mice, as inbred strains inherently lack the genetic diversity that would necessitate pangenome-based bioinformatic analysis. An additional forseeable difficulty arises when comparing transcriptomes across species as the problem of determining transcript equivalence would even further exacerbated based on newfound variation. Approaches have been developed using normal reference genomes to call the same transcripts and exons across species but this problem remains to be solved in pantranscriptomics[160,161]. Moreover, existing pantranscriptomic methods are designed exclusively for short-read RNA-seq and remain to be adapted and optimized for application to LR-RNA-seq data.

Optimizing computational and experimental protocols to improve long-read single-cell RNA-seq

The sparsity of single-cell RNA-seq data, coupled with the already-sparse nature of quantification of transcripts rather than genes as well as the difficulty in recovering full-length transcripts from single-cell or single-nucleus library preparation protocols, demonstrate that there still remains much to do to improve the quality and throughput of long-read single-cell RNA-seq experiments. Long-read single-nucleus experiments currently suffer from problems due to off-target capture of unspliced RNA from the nucleus and transcripts that do not appear to be full-length. Both our group and others have demonstrated that performing exon capture enrichment in single-nucleus LR-RNA-seq protocols immensely improves spliced read yield[76] (Fig. 4.1). Of course, this selection step is accompanied by a natural bias based on the capture panel being used, and additional work could be done to capture spliced RNA from the nucleus in a less biased manner. For example, since splicing is known to occur co-transcriptionally[162], the nucleoplasmic fraction of RNAs, which should be fully processed in comparison to chromatin-associated RNA, can be selectively amplified.

The abundance of seemingly non full-length transcripts could be addressed in several ways. To enrich for full-length transcripts rather than incomplete artifacts of reverse transcription, 5' end capture enrichment coupled with oligo dT priming could be done[62]. As previously mentioned, current combinatorial-based barcoding approaches[80,82] rely on a fixation step that may hinder full-length reverse transcription due to rigid secondary structures and bound proteins, which have been immobilized. To address this limitation, an alternative approach could involve reversing the crosslinks before performing reverse transcription on the cDNA. In the current methods, reverse transcription occurs during the barcoding phase using complementary sequence primers, which would not be compatible with this alternative approach. Instead, the barcoding strategy could employ blunt-end ligation of cell barcodes directly onto the RNA molecule, enabling the crosslinks to be reversed before reverse transcription. Implementing any of these strategies could enhance the yield of informative reads from long-read single-cell or nucleus RNA-seq experiments, addressing the sparsity issue and enabling more robust conclusions regarding cell type specificity of transcript isoform expression.

The experimental portions of long-read single-cell RNA-seq experiments are not the only aspect that long-read single-cell RNA-seq could be improved. Aside from demultiplexing long reads to their cells or nuclei of origin, there is a distinct lack of effort on improving how transcript-level information is used to perform classical single cell preprocessing tasks, such as feature selection, dimensionality reduction, and clustering, which influence how different cell types are called. As researchers care about both absolute transcript expression and relative transcript usage, it would follow that both of these numbers should be taken into account, and it is unclear now how this could be done given their inherent dependence. Thoughtful metrics that take both into account might be able to reveal cell subtypes or cell states that would be otherwise undiscoverable with short-read single-cell methods. Therefore, efforts should be made to improve the underlying statistical methods and metrics used to process long-read single-cell RNA-seq data.

Profiling the spatial specificity of transcript isoform expression

Many organs are known to have complex spatial organization and show regional-specific profiles of gene expression. For instance, the human brain cortex is divided into distinct cortical layers that each have characteristic patterns of gene expression[163], and the regional specificity of gene expression is known to be altered in certain disease conditions[164]. Spatial transcriptomic techniques employ a spatial barcoding strategy such that RNA molecules within each "spot" are all assigned the same barcode, allowing for region-resolved transcriptomic profiling. One study has used LR-RNA-seq in conjunction with spatial barcoding in the mouse brain. However, they did not examine full-length transcript isoforms and instead focused in specific AS events like exon skipping; suggesting that further improvements could be made to enable full-length capture of transcript isoforms[57].

Improving computational tools to analyze long-read RNA-seq data

Many computational tools that were developed to process and analyze LR-RNA-seq datasets even just a few years ago are now struggling to keep up with the orders of magnitude increase in throughput of long-read sequencing platforms. Though many more newly-developed tools are much faster[37], others are simply not feasible to cope with the raw number of reads that are now generated. Existing and new methods should be written to use multithreading capabilities as well as sparse and efficient data representations in order

Additionally, as the structure of samples used for a given study grow more complex, there is a need for statistical methods tools that can perform more complex statistical tests on LR-RNA-seq datasets. For instance, multi-way differential gene expression tests exist, but differential isoform usage methods are always limited to pairwise comparisons. Moreover, in datasets with continuous value metadata like timepoint, age, or disease progression score, it is logical to incorporate differential isoform usage tests that consider these factors. Similar to analogous methods used for differential gene expression testing[165], these approaches can unveil genes with isoform expression not only differing between two conditions but also displaying significant variability across multiple conditions. Similarly, for inquiries involving the coordination of alternative TSS, splicing, and TES usage, it is valuable to develop approaches that extend beyond computing pairwise coordination scores for these transcript features. Novel methods capable of detecting complex coordination events involving multiple exons would offer deeper insights into the mechanisms and coordination of transcription and transcript processing.

Network based approaches to identify regulators of transcript isoform identity

Network analysis is a common technique used in genomics and transcriptomics to describe relationships between genes. In gene regulatory networks (GRNs), directional links between genes imply regulatory relationships between genes where the source of an edge is the regulator gene and the sink is the target gene. GRNs are often used to describe how the expression of transcription factors influences target gene expression. Coexpression networks are able to infer sets of features (in most cases genes) that are coregulated across a set of samples. Though these relationships are informative to building GRNs, there is no directionality of coregulatory relationship in said networks. Application of network analysis methods would be well-suited to helping understand what the regulators driving transcript isoform choice are and how transcript isoform choice is coregulated. However, a grand challenge in this problem would be what features to include in such a network. Specific transcript isoforms of splicing factors and RNA binding proteins influence the splicing of their target transcripts in different ways[17]. Furthermore, as transcript isoform expression is definitionally dependent on the expression of the gene that it derives from, it is clear that representing these factors by their gene expression values alone would not be sufficient.

Assessing transcriptomic diversity across additional sources

In Chapter 3, I presented a way that we can assess and quantify transcript diversity between a set of transcripts based on the TSS, exon junction chain, and TES used in each transcript. However, these three transcript features are not the only defining characteristics of any given transcript. As mentioned earlier, transcripts can vary based on genetic variation, both in terms of SNP and SV content. Furthermore, transcripts can differ in RNA modifications. Over 150 unique varieties of RNA modifications, which happen post-transcriptionally, have been characterized[166,167]. Termed epitranscriptomics, this field holds promise to investigate how RNA modifications confer differential function to the end product of the transcript. Several diseases have been linked to disruption of genes that perform RNA editing changes[168] and therefore including RNA modifications as a level of information about transcript diversity would help understand the landscape of functional RNA product diversity. The ONT long-read platform is currently the only sequencing technique that is capable of detecting such modifications as it is able to sequence RNA directly (dRNA). This protocol has been particularly difficult to work with in the past due to its high input sample requirements, lower throughput, and specific error rate[169]. However, current reports from ONT claim that they have new chemistry versions for dRNA that will increase throughput 3-4x compared to existing chemistry, and improve error rates[170]. Therefore ONT's hypothetically infinitely-tunable basecalling, capacity to sequence dRNA, and promised improvements in throughput and error rate hold promise for being able to identify and read out these modifications in individual molecules at scale[167]. In the future, both genetic variation as well as epitranscriptomic characteristics of transcripts should be incorporated to characterize transcritpional diversity on a more resolved level that might yield additional mechanistic insights into differential RNA product function.

Bibliography

[1] Steven L. Salzberg. Open questions: How many genes do we have? *BMC Biology*, 16 (1):94, 2018. doi: 10.1186/s12915-018-0564-x.

[2] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigo, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale,

Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, 2001. ISSN 0036-8075. doi: 10.1126/science.1058040.

[3] International Human Genome Sequencing Consortium, Center for Genome Research:, Whitehead Institute for Biomedical Research, Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, The Sanger Centre:, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Washington University Genome Sequencing Center, Robert H Waterston, Richard K Wilson, LaDeana W Hillier, John D McPherson, Marco A Marra, Elaine R Mardis, Lucinda A Fulton, Asif T Chinwalla, Kymberlie H Pepin, Warren R Gish, Stephanie L Chissoe, Michael C Wendl, Kim D Delehaunty, Tracie L Miner, Andrew Delehaunty, Jason B Kramer, Lisa L Cook, Robert S Fulton, Douglas L Johnson, Patrick J Minx, Sandra W Clifton, US DOE Joint Genome Institute:, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Baylor College of Medicine Human Genome Sequencing Center:, Richard A Gibbs, Donna M Muzny, Steven E Scherer, John B Bouck, Erica J Sodergren, Kim C Worley, Catherine M Rives, James H Gorrell, Michael L Metzker, Susan L Naylor, Raju S Kucherlapati, David L Nelson, George M Weinstock, RIKEN Genomic Sciences Center:, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, UMR-8030:, Genoscope and CNRS, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, Institute of Molecular Biotechnology:, Department of Genome Analysis, André Rosen-

thal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, GTC Sequencing Center:, Douglas R Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Beijing Genomics Institute/Human Genome Center:, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, The Institute for Systems Biology:, Multimegabase Sequencing Center, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Stanford Genome Technology Center:, Ronald W Davis, Nancy A Federspiel, A Pia Abola, Michael J Proctor, University of Oklahoma's Advanced Center for Genome Technology:, Bruce A Roe, Feng Chen, Huaqin Pan, Max Planck Institute for Molecular Genetics:, Juliane Ramser, Hans Lehrach, Richard Reinhardt, Lita Annenberg Hazen Genome Center:, Cold Spring Harbor Laboratory, W Richard McCombie, Melissa de la Bastide, Neilay Dedhia, GBF—German Research Centre for Biotechnology:, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L Aravind, Jeffrey A Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G Brown, Christopher B Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R Copley, Tobias Doerks, Sean R Eddy, Evan E Eichler, Terrence S Furey, James Galagan, James G R Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L Steven Johnson, Thomas A Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W James Kent, Paul Kitts, Eugene V Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V Moran, Nicola Mulder, Victor J Pollara, Chris P Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F A Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I Wolf, Kenneth H Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, US National Institutes of Health:, Scientific management: National Human Genome Research Institute, Francis Collins, Mark S Guyer, Jane Peterson, Adam Felsenfeld, Kris A Wetterstrand, Stanford Human Genome Center:, Richard M Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R Cox, University of Washington Genome Center:, Maynard V Olson, Rajinder Kaul, Christopher Raymond, Keio University School of Medicine:, Department of Molecular Biology, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, University of Texas Southwestern Medical Center at Dallas:, Glen A Evans, Maria Athanasiou, Roger Schultz, US Department of Energy:, Office of Science, Aristides Patrinos, The Wellcome Trust:, and Michael J Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. ISSN 0028-0836. doi: 10.1038/35057062.

[4] Jonathan Hodgkin. What does a worm want with 20,000 genes? *Genome Biology*, 2 (11):comment2008.1, 2001. ISSN 1465-6906. doi: 10.1186/gb-2001-2-11-comment2008.

[5] Eddie Park, Zhicheng Pan, Zijun Zhang, Lan Lin, and Yi Xing. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *The American Journal of Human Genetics*, 102(1):11–26, 2018. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.11.002.

[6] Dafne Campigli Di Giammartino, Kensei Nishida, and James L. Manley. Mechanisms and Consequences of Alternative Polyadenylation. *Molecular Cell*, 43(6):853–866, 2011. ISSN 1097-2765. doi: 10.1016/j.molcel.2011.08.017.

[7] Takeshi Ara, Fabrice Lopez, William Ritchie, Philippe Benech, and Daniel Gautheret.

Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics*, 7(1):189, 2006. doi: 10.1186/1471-2164-7-189.

[8] Yi Xing and Christopher Lee. Alternative splicing and RNA selection pressure — evolutionary consequences for eukaryotic genomes. *Nature Reviews Genetics*, 7(7): 499–509, 2006. ISSN 1471-0056. doi: 10.1038/nrg1896.

[9] Hideki Nagasaki, Masanori Arita, Tatsuya Nishizawa, Makiko Suwa, and Osamu Gotoh. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene*, 364:53–62, 2005. ISSN 0378-1119. doi: 10.1016/j.gene.2005.07.027.

[10] Alan G. Hinnebusch, Ivaylo P. Ivanov, and Nahum Sonenberg. Translational control by 5-untranslated regions of eukaryotic mRNAs. *Science*, 352(6292):1413–1416, 2016. ISSN 0036-8075. doi: 10.1126/science.aad9868.

[11] Kelsey C. Martin and Anne Ephrussi. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell*, 136(4):719–730, 2009. ISSN 0092-8674. doi: 10.1016/j.cell.2009.01.044.

[12] Megan Stevens and Sebastian Oltean. Modulation of the Apoptosis Gene Bcl-x Function Through Alternative Splicing. *Frontiers in Genetics*, 10:804, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00804.

[13] Kian Huat Lim, Luciana Ferraris, Madeleine E. Filloux, Benjamin J. Raphael, and William G. Fairbrother. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences*, 108(27):11093–11098, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1101135108.

[14] Núria López-Bigas, Benjamin Audit, Christos Ouzounis, Genís Parra, and Roderic Guigó. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters*, 579(9):1900–1903, 2005. ISSN 0014-5793. doi: 10.1016/j.febslet.2005.02.047.

[15] Mirella Meregalli, Simona Maciotta, Valentina Angeloni, and Yvan Torrente. Duchenne muscular dystrophy caused by a frame-shift mutation in the acceptor splice site of intron 26. *BMC Medical Genetics*, 17(1):55, 2016. doi: 10.1186/s12881-016-0318-y.

[16] Kathryn R. Bowles, Derian A. Pugh, Laura-Maria Oja, Benjamin M. Jadow, Kurt Farrell, Kristen Whitney, Abhijeet Sharma, Jonathan D. Cherry, Towfique Raj, Ana C. Pereira, John F. Crary, and Alison M. Goate. Dysregulated coordination of MAPT exon 2 and exon 10 splicing underlies different tau pathologies in PSP and AD. *Acta Neuropathologica*, 143(2):225–243, 2022. ISSN 0001-6322. doi: 10.1007/s00401-021-02392-2.

[17] Chun-Hao Su, Dhananjaya D, and Woan-Yuh Tarn. Alternative Splicing in Neurogenesis and Brain Development. *Frontiers in Molecular Biosciences*, 5:12, 2018. ISSN 2296-889X. doi: 10.3389/fmolb.2018.00012.

[18] Christopher S. Bland, Eric T. Wang, Anthony Vu, Marjorie P. David, John C. Castle, Jason M. Johnson, Christopher B. Burge, and Thomas A. Cooper. Global regulation of alternative splicing during myogenic differentiation. *Nucleic Acids Research*, 38(21): 7651–7664, 2010. ISSN 0305-1048. doi: 10.1093/nar/gkq614.

[19] Soji Sebastian, Hervé Faralli, Zizhen Yao, Patricia Rakopoulos, Carmen Palii, Yi Cao, Kulwant Singh, Qi-Cai Liu, Alphonse Chu, Arif Aziz, Marjorie Brand, Stephen J. Tapscott, and F. Jeffrey Dilworth. Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation. *Genes & Development*, 27

(11):1247–1259, 2013. ISSN 0890-9369. doi: 10.1101/gad.215400.113.

[20] John Parkinson and Mark Blaxter. Expressed Sequence Tags (ESTs), Generation and Analysis. *Methods in Molecular Biology*, 533:1–12, 2009. ISSN 1064-3745. doi: 10.1007/978-1-60327-136-3_1.

[21] Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Carles Boix, Silvia Carbonell Sala, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Kevin L Howe, Toby Hunt, Osagie G Izuogu, Rory Johnson, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Ferriol Calvet Riera, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Maxim Y Wolf, Jinuri Xu, Yucheng T Yang, Andrew Yates, Daniel Zerbino, Yan Zhang, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Michael L Tress, and Paul Flicek. GENCODE 2021. *Nucleic Acids Research*, 49(D1):gkaa1087–, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1087.

[22] Adi L. Tarca, Roberto Romero, and Sorin Draghici. Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*, 195(2): 373–388, 2006. ISSN 0002-9378. doi: 10.1016/j.ajog.2006.07.001.

[23] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008. ISSN 1548-7091. doi: 10.1038/nmeth.1226.

[24] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J Heron, and Daniel J Turner. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206, 2018. ISSN 1548-7091. doi: 10.1038/nmeth.4577.

[25] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, 2009. ISSN 0036-8075. doi: 10.1126/science.1162986.

[26] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015. ISSN 1672-0229. doi: 10.1016/j.gpb.2015.08.002.

[27] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019. ISSN 1087-0156. doi: 10.1038/s41587-019-0217-9.

[28] Roger Volden, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard E. Green, and Christopher Vollmers. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences*, 115(39):9726–9731, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806447115.

[29] Oxford Nanopore. Accuracy, 2023. URL `https://nanoporetech.com/accuracy`.

[30] Aziz M. Al'Khafaji, Jonathan T. Smith, Kiran V. Garimella, Mehrtash Babadi, Victoria Popic, Moshe Sade-Feldman, Michael Gatzen, Siranush Sarkizova, Marc A. Schwartz, Emily M. Blaum, Allyson Day, Maura Costello, Tera Bowers, Stacey Gabriel, Eric Banks, Anthony A. Philippakis, Genevieve M. Boland, Paul C. Blainey, and Nir Hacohen. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nature Biotechnology*, pages 1–5, 2023. ISSN 1087-0156. doi: 10.1038/s41587-023-01815-7.

[31] Oxford Nanopore. Flow cells, 2023. URL `https://nanoporetech.com/how-it-works/flow-cells-and-nanopores`.

[32] Shanika L Amarasinghe, Matthew E Ritchie, and Quentin Gouil. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. *GigaScience*, 10(2):giab003, 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab003.

[33] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34 (18):3094–3100, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191.

[34] PacBio. Iso seq, 2023. URL `https://isoseq.how/`.

[35] Alison D. Tang, Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, 11(1):1438, 2020. doi: 10.1038/s41467-020-15171-6.

[36] Dana Wyman, Gabriela Balderrama-Gutierrez, Fairlie Reese, Shan Jiang, Sorena Rahmanian, Stefania Forner, Dina Matheos, Weihua Zeng, Brian Williams, Diane Trout, Whitney England, Shu-Hui Chu, Robert C. Spitale, Andrea J. Tenner, Barbara J. Wold, and Ali Mortazavi. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*, page 672931, 2020. doi: 10.1101/672931.

[37] Matthias Lienhard, Twan van den Beucken, Bernd Timmermann, Myriam Hochradel, Stefan Börno, Florian Caiment, Martin Vingron, and Ralf Herwig. IsoTools: a flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics*, 39(6): btad364, 2023. ISSN 1367-4803. doi: 10.1093/bioinformatics/btad364.

[38] Dana Wyman and Ali Mortazavi. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, 35(2):340–342, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty483.

[39] Inês Lopes, Gulam Altab, Priyanka Raina, and João Pedro de Magalhães. Gene Size Matters: An Analysis of Gene Length in the Human Genome. *Frontiers in Genetics*, 12:559998, 2021. ISSN 1664-8021. doi: 10.3389/fgene.2021.559998.

[40] Sam Kovaka, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, 20(1):278, 2019. ISSN 1474-7596. doi: 10.1186/s13059-019-1910-1.

[41] Muhammed Hasan Çelik and Ali Mortazavi. Analysis of alternative polyadenylation from long-read or short-read RNA-seq with LAPA. *bioRxiv*, page 2022.11.08.515683, 2022. doi: 10.1101/2022.11.08.515683.

[42] Manuel Tardaguila, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector del Risco, Marc Ferrell, Maravillas Mellado, Marissa Macchietto, Kenneth Verheggen, Mariola Edelmann, Iakes Ezkurdia, Jesus Vazquez, Michael Tress, Ali Mortazavi, Lennart Martens, Susana Rodriguez-Navarro, Victoria Moreno-Manzano, and Ana Conesa. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, 28(3):396–411, 2018. ISSN 1088-9051. doi: 10.1101/gr.222976.117.

[43] Andrey Prjibelski, Alla Mikheenko, Anoushka Joglekar, Alexander Smetanin, Julien Jarroux, Alla Lapidus, and Hagen Tilgner. IsoQuant: a tool for accurate novel isoform discovery with long reads. doi: 10.21203/rs.3.rs-1571850/v1.

[44] Angela Brooks, Francisco Pardo-Palacios, Fairlie Reese, Silvia Carbonell-Sala, Mark Diekhans, Cindy Liang, Dingjie Wang, Brian Williams, Matthew Adams, Amit Behera, Julien Lagarde, Haoran Li, Andrey Prjibelski, Gabriela Balderrama-Gutierrez, Muhammed Hasan Çelik, Maite De María, Nancy Denslow, Natàlia Garcia-Reyero, Stefan Goetz, Margaret Hunter, Jane Loveland, Carlos Menor, David Moraga, Jonathan Mudge, Hazuki Takahashi, Alison Tang, Ingrid Youngworth, Piero Carninci, Roderic Guigó, Hagen Tilgner, Barbara Wold, Christopher Vollmers, Gloria Sheynkman, Adam Frankish, Kin Fai Au, Ana Conesa, and Ali Mortazavi. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. 2021. doi: 10.21203/rs.3.rs-777702/v1.

[45] Yu Hu, Li Fang, Xuelian Chen, Jiang F. Zhong, Mingyao Li, and Kai Wang. LIQA: long-read isoform quantification and analysis. *Genome Biology*, 22(1):182, 2021. ISSN 1474-7596. doi: 10.1186/s13059-021-02399-8.

[46] Yuan Gao, Feng Wang, Robert Wang, Eric Kutschera, Yang Xu, Stephan Xie, Yuanyuan Wang, Kathryn E. Kadash-Edmondson, Lan Lin, and Yi Xing. ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Science Advances*, 9(3):eabq5072, 2023. doi: 10.1126/sciadv.abq5072.

[47] Ying Chen, Andre Sim, Yuk Kei Wan, Keith Yeo, Joseph Jing Xian Lee, Min Hao Ling, Michael I. Love, and Jonathan Göke. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nature Methods*, pages 1–9, 2023. ISSN 1548-7091. doi: 10.1038/s41592-023-01908-w.

[48] Baraa Orabi, Ning Xie, Brian McConeghy, Xuesen Dong, Cedric Chauve, and Faraz Hach. Freddie: annotation-independent detection and discovery of transcriptomic alternative splicing isoforms using long-read sequencing. *Nucleic Acids Research*, 51(2): e11–e11, 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1112.

[49] Jose Manuel Rodriguez, Paolo Maietta, Iakes Ezkurdia, Alessandro Pietrelli, Jan-Jaap Wesselink, Gonzalo Lopez, Alfonso Valencia, and Michael L. Tress. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, 41(D1): D110–D117, 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1058.

[50] Joannella Morales, Shashikant Pujar, Jane E. Loveland, Alex Astashyn, Ruth Bennett, Andrew Berry, Eric Cox, Claire Davidson, Olga Ermolaeva, Catherine M. Farrell, Reham Fatima, Laurent Gil, Tamara Goldfarb, Jose M. Gonzalez, Diana Haddad, Matthew Hardy, Toby Hunt, John Jackson, Vinita S. Joardar, Michael Kay, Vamsi K. Kodali, Kelly M. McGarvey, Aoife McMahon, Jonathan M. Mudge, Daniel N. Murphy, Michael R. Murphy, Bhanu Rajput, Sanjida H. Rangwala, Lillian D. Riddick, Françoise Thibaud-Nissen, Glen Threadgold, Anjana R. Vatsan, Craig Wallin, David Webb, Paul Flicek, Ewan Birney, Kim D. Pruitt, Adam Frankish, Fiona Cunningham, and Terence D. Murphy. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604(7905):310–315, 2022. ISSN 0028-0836. doi: 10.1038/s41586-022-04558-8.

[51] Richard I. Kuo, Yuanyuan Cheng, Runxuan Zhang, John W. S. Brown, Jacqueline Smith, Alan L. Archibald, and David W. Burt. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*, 21(1): 751, 2020. doi: 10.1186/s12864-020-07123-7.

[52] Lorena de la Fuente, Ángeles Arzalluz-Luque, Manuel Tardáguila, Héctor del Risco, Cristina Martí, Sonia Tarazona, Pedro Salguero, Raymond Scott, Alberto Lerma, Ana Alastrue-Agudo, Pablo Bonilla, Jeremy R. B. Newman, Shunichi Kosugi, Lauren M. McIntyre, Victoria Moreno-Manzano, and Ana Conesa. tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biology*, 21(1):119, 2020. ISSN 1474-7596. doi: 10.1186/s13059-020-02028-w.

[53] Rachel M. Miller, Ben T. Jordan, Madison M. Mehlferber, Erin D. Jeffery, Christina Chatzipantsiou, Simi Kaur, Robert J. Millikin, Yunxiang Dai, Simone Tiberi, Peter J. Castaldi, Michael R. Shortreed, Chance John Luckey, Ana Conesa, Lloyd M. Smith, Anne Deslattes Mays, and Gloria M. Sheynkman. Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biology*, 23(1):69, 2022. ISSN 1474-7596. doi: 10.1186/s13059-022-02624-y.

[54] Liang Niu, Weichun Huang, David M Umbach, and Leping Li. IUTA: a tool for effectively detecting differential isoform usage from RNA-Seq data. *BMC Genomics*, 15(1):862, 2014. doi: 10.1186/1471-2164-15-862.

[55] Kristoffer Vitting-Seerup and Albin Sandelin. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, 35(21):4469–4471, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz247.

[56] Fairlie Reese and Ali Mortazavi. Swan: a library for the analysis and visualization of long-read transcriptomes. *Bioinformatics*, 37(9):btaa836–, 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa836.

[57] Anoushka Joglekar, Andrey Prjibelski, Ahmed Mahfouz, Paul Collier, Susan Lin, Anna Katharina Schlusche, Jordan Marrocco, Stephen R. Williams, Bettina Haase, Ashley Hayes, Jennifer G. Chew, Neil I. Weisenfeld, Man Ying Wong, Alexander N. Stein, Simon A. Hardwick, Toby Hunt, Qi Wang, Christoph Dieterich, Zachary Bent, Olivier Fedrigo, Steven A. Sloan, Davide Risso, Erich D. Jarvis, Paul Flicek, Wenjie Luo, Geoffrey S. Pitt, Adam Frankish, August B. Smit, M. Elizabeth Ross, and Hagen U. Tilgner. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nature Communications*, 12(1):463, 2021. doi: 10.1038/s41467-020-20343-5.

[58] Beril Erdogdu, Ales Varabyou, Stephanie C Hicks, Steven L Salzberg, and Mihaela Pertea. Detecting differential transcript usage in complex diseases with SPIT. 2023. doi: 10.1101/2023.07.10.548289.

[59] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010. ISSN 1548-7091. doi: 10.1038/nmeth.1528.

[60] Alexander N Stein, Anoushka Joglekar, Chi-Lam Poon, and Hagen U Tilgner. Scisor-Wiz: Visualizing Differential Isoform Expression in Single-Cell Long-Read Data. *Bioinformatics*, 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac340.

[61] Emil K Gustavsson, David Zhang, Regina H Reynolds, Sonia Garcia-Ruiz, and Mina Ryten. ggtranscript: an R package for the visualization and interpretation of transcript isoforms using ggplot2. *Bioinformatics*, 38(15):3844–3846, 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac409.

[62] Silvia Carbonell-Sala, Julien Lagarde, Hiromi Nishiyori, Emilio Palumbo, Carme Arnan, Hazuki Takahashi, Piero Carninci, Barbara Uszczynska-Ratajczak, and Roderic Guigó. CapTrap-Seq: A platform-agnostic and quantitative approach for high-fidelity full-length RNA transcript sequencing. *bioRxiv*, page 2023.06.16.543444, 2023. doi: 10.1101/2023.06.16.543444.

[63] Tony Kwan, David Benovoy, Christel Dias, Scott Gurd, Cathy Provencher, Patrick Beaulieu, Thomas J Hudson, Rob Sladek, and Jacek Majewski. Genome-wide analysis of transcript isoform variation in humans. *Nature Genetics*, 40(2):225–231, 2008. ISSN 1061-4036. doi: 10.1038/ng.2007.57.

[64] Raquel García-Pérez, Jose Miguel Ramirez, Aida Ripoll-Cladellas, Ruben Chazarra-Gil, Winona Oliveros, Oleksandra Soldatkina, Mattia Bosio, Paul Joris Rognon, Salvador Capella-Gutierrez, Miquel Calvo, Ferran Reverter, Roderic Guigó, François Aguet, Pedro G. Ferreira, Kristin G. Ardlie, and Marta Melé. The landscape of expression and alternative splicing variation across human traits. *Cell Genomics*, 3(1): 100244, 2023. ISSN 2666-979X. doi: 10.1016/j.xgen.2022.100244.

[65] Eugene Melamud and John Moult. Stochastic noise in splicing machinery. *Nucleic Acids Research*, 37(14):4873–4886, 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp471.

[66] Joseph K. Pickrell, Athma A. Pai, Yoav Gilad, and Jonathan K. Pritchard. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genetics*, 6(12):e1001236, 2010. ISSN 1553-7390. doi: 10.1371/journal.pgen.1001236.

[67] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008. ISSN 1061-4036.

doi: 10.1038/ng.259.

[68] Barmak Modrek and Christopher J Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics*, 34(2):177–180, 2003. ISSN 1061-4036. doi: 10.1038/ng1159.

[69] Serafim Batzoglou, Lior Pachter, Jill P. Mesirov, Bonnie Berger, and Eric S. Lander. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. *Genome Research*, 10(7):950–958, 2000. ISSN 1088-9051. doi: 10.1101/gr.10.7.950.

[70] Lan Lin, Peng Jiang, Juw Won Park, Jinkai Wang, Zhi-xiang Lu, Maggie P. Y. Lam, Peipei Ping, and Yi Xing. The contribution of Alu exons to the human proteome. *Genome Biology*, 17(1):15, 2016. ISSN 1474-7596. doi: 10.1186/s13059-016-0876-5.

[71] Mar Gonzàlez-Porta, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14(7):R70, 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-7-r70.

[72] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, 2004. ISSN 0036-8075. doi: 10.1126/science.1105136.

[73] Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012. ISSN 0028-0836. doi: 10.1038/nature11233.

[74] Alessandra Breschi, Manuel Muñoz-Aguirre, Valentin Wucher, Carrie A. Davis, Diego Garrido-Martín, Sarah Djebali, Jesse Gillis, Dmitri D. Pervouchine, Anna Vlasova, Alexander Dobin, Chris Zaleski, Jorg Drenkow, Cassidy Danyko, Alexandra Scavelli, Ferran Reverter, Michael P. Snyder, Thomas R. Gingeras, and Roderic Guigó. A limited set of transcriptional programs define major cell types. *Genome Research*, 30 (7):1047–1059, 2020. ISSN 1088-9051. doi: 10.1101/gr.263186.120.

[75] Dafni A. Glinos, Garrett Garborcauskas, Paul Hoffman, Nava Ehsan, Lihua Jiang, Alper Gokden, Xiaoguang Dai, François Aguet, Kathleen L. Brown, Kiran Garimella, Tera Bowers, Maura Costello, Kristin Ardlie, Ruiqi Jian, Nathan R. Tucker, Patrick T.

Ellinor, Eoghan D. Harrington, Hua Tang, Michael Snyder, Sissel Juul, Pejman Mohammadi, Daniel G. MacArthur, Tuuli Lappalainen, and Beryl B. Cummings. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature*, pages 1–8, 2022. ISSN 0028-0836. doi: 10.1038/s41586-022-05035-y.

[76] Simon A. Hardwick, Wen Hu, Anoushka Joglekar, Li Fan, Paul G. Collier, Careen Foord, Jennifer Balacco, Samantha Lanjewar, Maureen McGuirk Sampson, Frank Koopmans, Andrey D. Prjibelski, Alla Mikheenko, Natan Belchikov, Julien Jarroux, Anne Bergstrom Lucas, Miklós Palkovits, Wenjie Luo, Teresa A. Milner, Lishomwa C. Ndhlovu, August B. Smit, John Q. Trojanowski, Virginia M. Y. Lee, Olivier Fedrigo, Steven A. Sloan, Dóra Tombácz, M. Elizabeth Ross, Erich Jarvis, Zsolt Boldogkői, Li Gan, and Hagen U. Tilgner. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nature Biotechnology*, 40(7):1082–1092, 2022. ISSN 1087-0156. doi: 10.1038/s41587-022-01231-3.

[77] Kirsten A. Reimer, Claudia A. Mimoso, Karen Adelman, and Karla M. Neugebauer. Co-transcriptional splicing regulates 3 end cleavage during mammalian erythropoiesis. *Molecular Cell*, 81(5):998–1012.e7, 2021. ISSN 1097-2765. doi: 10.1016/j.molcel.2020.12.018.

[78] Carlos Alfonso-Gonzalez, Ivano Legnini, Sarah Holec, Laura Arrigoni, Hasan Can Ozbulut, Fernando Mateos, David Koppstein, Agnieszka Rybak-Wolf, Ulrike Bönisch, Nikolaus Rajewsky, and Valérie Hilgers. Sites of transcription initiation drive mRNA isoform selection. *Cell*, 2023. ISSN 0092-8674. doi: 10.1016/j.cell.2023.04.012.

[79] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017. doi: 10.1038/ncomms14049.

[80] Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360 (6385):176–182, 2018. ISSN 0036-8075. doi: 10.1126/science.aam8999.

[81] Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4):1103–1116.e20, 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.09.056.

[82] Beth K. Martin, Chengxiang Qiu, Eva Nichols, Melissa Phung, Rula Green-Gladden, Sanjay Srivatsan, Ronnie Blecher-Gonen, Brian J. Beliveau, Cole Trapnell, Junyue Cao, and Jay Shendure. Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nature Protocols*, 18(1):188–207, 2023. ISSN 1754-2189. doi: 10.1038/s41596-022-00752-0.

[83] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8):1–14, 2018. ISSN 1226-3613. doi: 10.1038/s12276-018-0071-8.

[84] Shan Jiang, Katherine Williams, Xiangduo Kong, Weihua Zeng, Nam Viet Nguyen, Xinyi Ma, Rabi Tawil, Kyoko Yokomori, and Ali Mortazavi. Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. *PLoS Genetics*, 16(5): e1008754, 2020. ISSN 1553-7390. doi: 10.1371/journal.pgen.1008754.

[85] Samuel Morabito, Emily Miyoshi, Neethu Michael, Saba Shahin, Alessandra Cadete Martini, Elizabeth Head, Justine Silva, Kelsey Leavy, Mari Perez-Rosendahl, and Vivek Swarup. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nature Genetics*, 53(8):1143–1155, 2021. ISSN 1061-4036. doi: 10.1038/s41588-021-00894-z.

[86] Alex K. Shalek, Rahul Satija, Xian Adiconis, Rona S. Gertner, Jellert T. Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J. Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z. Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, 2013. ISSN 0028-0836. doi: 10.1038/nature12172.

[87] Carlos F. Buen Abad Najar, Prakruthi Burra, Nir Yosef, and Liana F. Lareau. Identifying cell state–associated alternative splicing events and their coregulation. *Genome Research*, 32(7):1385–1397, 2022. ISSN 1088-9051. doi: 10.1101/gr.276109.121.

[88] Yuanhua Huang and Guido Sanguinetti. BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome Biology*, 22(1): 251, 2021. ISSN 1474-7596. doi: 10.1186/s13059-021-02461-5.

[89] Georgi K. Marinov, Brian A. Williams, Ken McCue, Gary P. Schroth, Jason Gertz, Richard M. Myers, and Barbara J. Wold. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, 2014. ISSN 1088-9051. doi: 10.1101/gr.161034.113.

[90] Carlos F Buen Abad Najar, Nir Yosef, and Liana F Lareau. Coverage-dependent bias creates the appearance of binary splicing in single cells. *eLife*, 9:e54603, 2020. doi: 10.7554/elife.54603.

[91] Anoushka Joglekar, Careen Foord, Julien Jarroux, Shaun Pollard, and Hagen U Tilgner. From words to complete phrases: insight into single-cell isoforms using short and long reads. *Transcription*, ahead-of-print(ahead-of-print):1–13, 2023. ISSN 2154-1264. doi: 10.1080/21541264.2023.2213514.

[92] Arthur Dondi, Ulrike Lischetti, Francis Jacob, Franziska Singer, Nico Borgsmüller, Tumor Profiler Consortium, Viola Heinzelmann-Schwarz, Christian Beisel, and Niko Beerenwinkel. Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing for personalized oncology. *bioRxiv*, page 2022.12.12.520051, 2022. doi: 10.1101/2022.12.12.520051.

[93] Carter R. Palmer, Christine S. Liu, William J. Romanow, Ming-Hsiang Lee, and Jerold Chun. Altered cell and RNA isoform diversity in aging Down syndrome brains. *Proceedings of the National Academy of Sciences*, 118(47):e2114326118, 2021. ISSN 0027-8424. doi: 10.1073/pnas.2114326118.

[94] Anoushka Joglekar, Wen Hu, Bei Zhang, Oleksandr Narykov, Mark Diekhans, Jennifer

Balacco, Lishomwa C Ndhlovu, Teresa A Milner, Olivier Fedrigo, Erich D Jarvis, Gloria Sheynkman, Dmitry Korkin, M. Elizabeth Ross, and Hagen U. Tilgner. Single-cell long-read mRNA isoform regulation is pervasive across mammalian brain regions, cell types, and development. *bioRxiv*, page 2023.04.02.535281, 2023. doi: 10.1101/2023.04.02.535281.

[95] Fairlie Reese. Lr-splitpipe, 2021. URL `https://github.com/fairliereese/LR-splitpipe`.

[96] Elisabeth Rebboah, Fairlie Reese, Katherine Williams, Gabriela Balderrama-Gutierrez, Cassandra McGill, Diane Trout, Isaryhia Rodriguez, Heidi Liang, Barbara J. Wold, and Ali Mortazavi. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. *Genome Biology*, 22(1): 286, 2021. doi: 10.1186/s13059-021-02505-w.

[97] Haojia Wu, Yuhei Kirita, Erinn L. Donnelly, and Benjamin D. Humphreys. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *Journal of the American Society of Nephrology*, 30(1):23–32, 2019. ISSN 1046-6673. doi: 10.1681/asn.2018090912.

[98] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 2016. ISSN 1474-7596. doi: 10.1186/s13059-016-0881-8.

[99] Yarden Katz, Eric T. Wang, Jacob Silterra, Schraga Schwartz, Bang Wong, Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov, Edoardo M. Airoldi, and Christopher B. Burge. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, 31(14):2400–2402, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv034.

[100] Yang I. Li, David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, 2018. ISSN 1061-4036. doi: 10.1038/s41588-017-0004-9.

[101] Fairlie Reese, Brian Williams, Gabriela Balderrama-Gutierrez, Dana Wyman, Muhammed Hasan Çelik, Elisabeth Rebboah, Narges Rezaie, Diane Trout, Milad Razavi-Mohseni, Yunzhe Jiang, Beatrice Borsari, Samuel Morabito, Heidi Yahan Liang, Cassandra J. McGill, Sorena Rahmanian, Jasmine Sakr, Shan Jiang, Weihua Zeng, Klebea Carvalho, Annika K. Weimer, Louise A. Dionne, Ariel McShane, Karan Bedi, Shaimae I. Elhajjajy, Sean Upchurch, Jennifer Jou, Ingrid Youngworth, Idan Gabdank, Paul Sud, Otto Jolanki, J. Seth Strattan, Meenakshi S. Kagda, Michael P. Snyder, Ben C. Hitz, Jill E. Moore, Zhiping Weng, David Bennett, Laura Reinholdt, Mats Ljungman, Michael A. Beer, Mark B. Gerstein, Lior Pachter, Roderic Guigó, Barbara J. Wold, and Ali Mortazavi. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *bioRxiv*, page 2023.05.15.540865, 2023. doi: 10.1101/2023.05.15.540865.

[102] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. ISSN 1474-7596. doi: 10.1186/s13059-017-1382-0.

[103] Fairlie Reese, Brian Williams, Gabriela Balderrama-Gutierrez, Dana Wyman, Muhammed Hasan Çelik, Elisabeth Rebboah, Narges Rezaie, Diane Trout, Milad Razavi-Mohseni, Yunzhe Jiang, Beatrice Borsari, Samuel Morabito, Heidi Yahan Liang, Cassandra J. McGill, Sorena Rahmanian, Jasmine Sakr, Shan Jiang, Weihua Zeng, Klebea Carvalho, Annika K. Weimer, Louise A. Dionne, Ariel McShane, Karan Bedi, , Shaimae I. Elhajjajy, Sean Upchurch, Jennifer Jou, Ingrid Youngworth, Idan Gabdank, Paul Sud, Otto Jolanki, J. Seth Strattan, Meenakshi S. Kagda, Michael P. Snyder, Ben C. Hitz, Jill E. Moore, Zhiping Weng, David Bennett, Laura Reinholdt, Mats Ljungman, Michael A. Beer, Mark B. Gerstein, Lior Pachter, Roderic Guigó, Barbara J. Wold, and Ali Mortazavi. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. 2023.

[104] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, 2012. ISSN 1088-9051. doi: 10.1101/gr.135350.111.

[105] Borwyn Wang and Hrishikesh Mehta. Cytokine receptor splice variants in hematologic diseases. *Cytokine*, 127:154919, 2020. ISSN 1043-4666. doi: 10.1016/j.cyto.2019.154919.

[106] Banglian Hu, Shengshun Duan, Ziwei Wang, Xin Li, Yuhang Zhou, Xian Zhang, Yun-Wu Zhang, Huaxi Xu, and Honghua Zheng. Insights Into the Role of CSF1R in the Central Nervous System and Neurological Disorders. *Frontiers in Aging Neuroscience*, 13:789834, 2021. ISSN 1663-4365. doi: 10.3389/fnagi.2021.789834.

[107] Elizabeth E Blue, Joshua C Bis, Michael O Dorschner, Debby W Tsuang, Sandra M Barral, Gary Beecham, Jennifer E Below, William S Bush, Mariusz Butkiewicz, Carlos Cruchaga, Anita DeStefano, Lindsay A Farrer, Alison Goate, Jonathan Haines, Jim Jaworski, Gyungah Jun, Brian Kunkle, Amanda Kuzma, Jenny J Lee, Kathryn L Lunetta, Yiyi Ma, Eden Martin, Adam Naj, Alejandro Q Nato, Patrick Navas, Hiep Nguyen, Christiane Reitz, Dolly Reyes, William Salerno, Gerard D Schellenberg, Sudha Seshadri, Harkirat Sohi, Timothy A Thornton, Otto Valadares, Cornelia van Duijn, Badri N Vardarajan, Li-San Wang, Eric Boerwinkle, Josée Dupuis, Margaret A Pericak-Vance, Richard Mayeux, Ellen M Wijsman, and on behalf of the Alzheimer's Disease Sequencing Project. Genetic Variation in Genes Underlying Diverse Dementias May Explain a Small Proportion of Cases in the Alzheimer's Disease Sequencing Project. *Dementia and Geriatric Cognitive Disorders*, 45(1-2):1–17, 2018. ISSN 1420-8008. doi: 10.1159/000485503.

[108] Violeta Chitu, Fabrizio Biundo, Gabriel G.L. Shlager, Eun S. Park, Ping Wang, Maria E. Gulinello, Şölen Gokhan, Harmony C. Ketchum, Kusumika Saha, Michael A. DeTure, Dennis W. Dickson, Zbignew K. Wszolek, Deyou Zheng, Andrew L. Croxford,

Burkhard Becher, Daqian Sun, Mark F. Mehler, and E. Richard Stanley. Microglial Homeostasis Requires Balanced CSF-1/CSF-2 Receptor Signaling. *Cell Reports*, 30 (9):3004–3019.e5, 2020. ISSN 2211-1247. doi: 10.1016/j.celrep.2020.02.028.

[109] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *bioRxiv*, page 2022.12.14.520412, 2022. doi: 10.1101/2022.12.14.520412.

[110] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1): D733–D745, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1189.

[111] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–19, 2002. ISSN 1061-4036. doi: 10.1038/ng0102-13.

[112] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Morgan C. Giddings, Bradley E. Bern-

stein, Charles B. Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D. Ward, Robert C. Altshuler, Matthew L. Eaton, Manolis Kellis, Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P. Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Brian A. Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J. Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R. Gingeras, Kate R. Rosenbloom, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, W. James Kent, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Terrence S. Furey, Lingyun Song, Linda L. Grasfeder, Paul G. Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C. Sheffield, Jeremy M. Simon, Kimberly A. Showers, Alexias Safi, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R. Iyer, Jason D. Lieb, Gregory E. Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J. Luo, Atif Shahab, Melissa J. Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M. Myers, Florencia Pauli, Brian A. Williams, Jason Gertz, Georgi K. Marinov, Timothy E. Reddy, Jost Vielmetter, E. Partridge, Diane Trout, Katherine E. Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M. Bowling, Michael Anaya, Marie K. Cross, Brandon King, Michael A. Muratet, Igor Antoshechkin, Kimberly M. Newberry, Kenneth McCue, Amy S. Nesmith, Katherine I. Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L. Parker, Sreeram Balasubramanian, Nicholas S. Davis, Sarah K. Meadows, Tracy Eggleston, Chris Gunter, J. Scott Newberry, Shawn E. Levy, Devin M. Absher, Ali Mortazavi, Wing H. Wong, Barbara Wold, Matthew J. Blow, Axel Visel, Len A. Pennachio, Laura Elnitski, Elliott H. Margulies, Stephen C. J. Parker, Hanna M. Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Jacqueline Chrast, Claire Davidson, Thomas Derrien, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A. Hendrix, Cédric Howald, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Felix Kokocinski, Jing Leng, Michael F. Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Andrea Tanzer, Electra Tapanari, Michael L. Tress, Marijke J. van Baren, Nathalie Walters, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhengdong Zhang, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Mark Gerstein, Alexandre Reymond, Roderic Guigó, Jennifer Harrow, Timothy J. Hubbard, Stephen G. Landt, Seth Frietze, Alexej Abyzov, Nick Addleman, Roger P. Alexander, Raymond K. Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P. Boyle, Alina R. Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Chao Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D. Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X. Jin, Konrad J. Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jing Leng, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J. Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Baikang Pei, Debasish Raha, Lucia Ramirez, Brian Reed, Joel Rozowsky, Andrea Sboner, Minyi Shi, Cristina Sisu, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon-Kiu Yan, Xinqiong Yang, Kevin Y. Yip, Zhengdong Zhang, Kevin Struhl, Sherman M. Weissman, Mark Gerstein, Peggy J. Farnham, Michael Snyder, Scott A. Tenenbaum, Luiz O. Penalva, Francis Doyle, Subhradip Karmakar, Stephen G. Landt, Raj R. Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Dorrelyn Patacsil, Teri Slifer, Alec Victorsen, Xinqiong Yang, Michael Snyder, Kevin P. White, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Zhiping Weng, Troy W. Whitfield, Jie Wang, Patrick J. Collins, Shelley F. Aldred, Nathan D. Trinklein, E. Christopher Partridge, Richard M. Myers, Job Dekker, Gaurav Jain, Bryan R. Lajoie, Amartya Sanyal, Gayathri Balasundaram, Daniel L. Bates, Rachel Byron, Theresa K. Canfield, Morgan J. Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R. Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Gaurav Jain, Audra K. Johnson, Ericka M. Johnson, Tattyana V. Kutyavin, Bryan R. Lajoie, Kristen Lee, Dimitra Lotakis, Matthew T. Maurano, Shane J. Neph, Fiedencio V. Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Eric Rynes, Peter Sabo, Minerva E. Sanchez, Richard S. Sandstrom, Amartya Sanyal, Anthony O. Shafer, Andrew B. Stergachis, Sean Thomas, Robert E. Thurman, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Hao Wang, Molly A. Weaver, Yongqi Yan, Miaohua Zhang, Joshua M. Akey, Michael Bender, Michael O. Dorschner, Mark Groudine, Michael J. MacCoss, Patrick Navas, George Stamatoyannopoulos, Rajinder Kaul, Job Dekker, John A. Stamatoyannopoulos, Ian Dunham, Kathryn Beal, Alvis Brazma, Paul Flicek, Javier Herrero, Nathan Johnson, Damian Keefe, Margus Lukk, Nicholas M. Luscombe, Daniel Sobral, Juan M. Vaquerizas, Steven P. Wilder, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia

Kyriazopoulou-Panagiotopoulou, Max W. Libbrecht, Marc A. Schaub, Anshul Kundaje, Ross C. Hardison, Webb Miller, Belinda Giardine, Robert S. Harris, Weisheng Wu, Peter J. Bickel, Balazs Banfai, Nathan P. Boley, James B. Brown, Haiyan Huang, Qunhua Li, Jingyi Jessica Li, William Stafford Noble, Jeffrey A. Bilmes, Orion J. Buske, Michael M. Hoffman, Avinash D. Sahu, Peter V. Kharchenko, Peter J. Park, Dannon Baker, James Taylor, Zhiping Weng, Sowmya Iyer, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Hualin S. Xi, Jiali Zhuang, Mark Gerstein, Roger P. Alexander, Suganthi Balasubramanian, Chao Cheng, Arif Harmanci, Lucas Lochovsky, Renqiang Min, Xinmeng J. Mu, Joel Rozowsky, Koon-Kiu Yan, Kevin Y. Yip, and Ewan Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. ISSN 0028-0836. doi: 10.1038/nature11247.

[113] Federico Abascal, Reyes Acosta, Nicholas J. Addleman, Jessika Adrian, Veena Afzal, Rizi Ai, Bronwen Aken, Jennifer A. Akiyama, Omar Al Jammal, Henry Amrhein, Stacie M. Anderson, Gregory R. Andrews, Igor Antoshechkin, Kristin G. Ardlie, Joel Armstrong, Matthew Astley, Budhaditya Banerjee, Amira A. Barkal, If H. A. Barnes, Iros Barozzi, Daniel Barrell, Gemma Barson, Daniel Bates, Ulugbek K. Baymuradov, Cassandra Bazile, Michael A. Beer, Samantha Beik, M. A. Bender, Ruth Bennett, Louis Philip Benoit Bouvrette, Bradley E. Bernstein, Andrew Berry, Anand Bhaskar, Alexandra Bignell, Steven M. Blue, David M. Bodine, Carles Boix, Nathan Boley, Tyler Borrman, Beatrice Borsari, Alan P. Boyle, Laurel A. Brandsmeier, Alessandra Breschi, Emery H. Bresnick, Jason A. Brooks, Michael Buckley, Christopher B. Burge, Rachel Byron, Eileen Cahill, Lingling Cai, Lulu Cao, Mark Carty, Rosa G. Castanon, Andres Castillo, Hassan Chaib, Esther T. Chan, Daniel R. Chee, Sora Chee, Hao Chen, Huaming Chen, Jia-Yu Chen, Songjie Chen, J. Michael Cherry, Surya B. Chhetri, Jyoti S. Choudhary, Jacqueline Chrast, Dongjun Chung, Declan Clarke, Neal A. L. Cody, Candice J. Coppola, Julie Coursen, Anthony M. D'Ippolito, Stephen Dalton, Cassidy Danyko, Claire Davidson, Jose Davila-Velderrain, Carrie A. Davis, Job Dekker, Alden Deran, Gilberto DeSalvo, Gloria Despacio-Reyes, Colin N. Dewey, Diane E. Dickel, Morgan Diegel, Mark Diekhans, Vishnu Dileep, Bo Ding, Sarah Djebali, Alexander Dobin, Daniel Dominguez, Sarah Donaldson, Jorg Drenkow, Timothy R. Dreszer, Yotam Drier, Michael O. Duff, Douglass Dunn, Catharine Eastman, Joseph R. Ecker, Matthew D. Edwards, Nicole El-Ali, Shaimae I. Elhajjajy, Keri Elkins, Andrew Emili, Charles B. Epstein, Rachel C. Evans, Iakes Ezkurdia, Kaili Fan, Peggy J. Farnham, Nina P. Farrell, Elise A. Feingold, Anne-Maud Ferreira, Katherine Fisher-Aylor, Stephen Fitzgerald, Paul Flicek, Chuan Sheng Foo, Kevin Fortier, Adam Frankish, Peter Freese, Shaliu Fu, Xiang-Dong Fu, Yu Fu, Yoko Fukuda-Yuzawa, Mariateresa Fulciniti, Alister P. W. Funnell, Idan Gabdank, Timur Galeev, Mingshi Gao, Carlos Garcia Giron, Tyler H. Garvin, Chelsea Anne Gelboin-Burkhart, Grigorios Georgolopoulos, Mark B. Gerstein, Belinda M. Giardine, David K. Gifford, David M. Gilbert, Daniel A. Gilchrist, Shawn Gillespie, Thomas R. Gingeras, Peng Gong, Alvaro Gonzalez, Jose M. Gonzalez, Peter Good, Alon Goren, David U. Gorkin, Brenton R. Graveley, Michael Gray, Jack F. Greenblatt, Ed Griffiths, Mark T. Groudine, Fabian Grubert, Mengting Gu, Roderic Guigó, Hongbo Guo, Yu Guo, Yuchun Guo, Gamze Gursoy, Maria Gutierrez-Arcelus, Jessica Halow, Ross C. Hardison, Matthew Hardy, Manoj Hariharan, Arif Harmanci, Anne Harrington, Jennifer L. Harrow, Tatsunori B. Hashimoto,

Richard D. Hasz, Meital Hatan, Eric Haugen, James E. Hayes, Peng He, Yupeng He, Nastaran Heidari, David Hendrickson, Elisabeth F. Heuston, Jason A. Hilton, Benjamin C. Hitz, Abigail Hochman, Cory Holgren, Lei Hou, Shuyu Hou, Yun-Hua E. Hsiao, Shanna Hsu, Hui Huang, Tim J. Hubbard, Jack Huey, Timothy R. Hughes, Toby Hunt, Sean Ibarrientos, Robbyn Issner, Mineo Iwata, Osagie Izuogu, Tommi Jaakkola, Nader Jameel, Camden Jansen, Lixia Jiang, Peng Jiang, Audra Johnson, Rory Johnson, Irwin Jungreis, Madhura Kadaba, Maya Kasowski, Mary Kasparian, Momoe Kato, Rajinder Kaul, Trupti Kawli, Michael Kay, Judith C. Keen, Sunduz Keles, Cheryl A. Keller, David Kelley, Manolis Kellis, Pouya Kheradpour, Daniel Sunwook Kim, Anthony Kirilusha, Robert J. Klein, Birgit Knoechel, Samantha Kuan, Michael J. Kulik, Sushant Kumar, Anshul Kundaje, Tanya Kutyavin, Julien Lagarde, Bryan R. Lajoie, Nicole J. Lambert, John Lazar, Ah Young Lee, Donghoon Lee, Elizabeth Lee, Jin Wook Lee, Kristen Lee, Christina S. Leslie, Shawn Levy, Bin Li, Hairi Li, Nan Li, Shantao Li, Xiangrui Li, Yang I. Li, Ying Li, Yining Li, Yue Li, Jin Lian, Maxwell W. Libbrecht, Shin Lin, Yiing Lin, Dianbo Liu, Jason Liu, Peng Liu, Tingting Liu, X. Shirley Liu, Yan Liu, Yaping Liu, Maria Long, Shaoke Lou, Jane Loveland, Aiping Lu, Yuheng Lu, Eric Lécuyer, Lijia Ma, Mark Mackiewicz, Brandon J. Mannion, Michael Mannstadt, Deepa Manthravadi, Georgi K. Marinov, Fergal J. Martin, Eugenio Mattei, Kenneth McCue, Megan McEown, Graham McVicker, Sarah K. Meadows, Alex Meissner, Eric M. Mendenhall, Christopher L. Messer, Wouter Meuleman, Clifford Meyer, Steve Miller, Matthew G. Milton, Tejaswini Mishra, Dianna E. Moore, Helen M. Moore, Jill E. Moore, Samuel H. Moore, Jennifer Moran, Ali Mortazavi, Jonathan M. Mudge, Nikhil Munshi, Rabi Murad, Richard M. Myers, Vivek Nandakumar, Preetha Nandi, Anil M. Narasimha, Aditi K. Narayanan, Hannah Naughton, Fabio C. P. Navarro, Patrick Navas, Jurijs Nazarovs, Jemma Nelson, Shane Neph, Fidencio Jun Neri, Joseph R. Nery, Amy R. Nesmith, J. Scott Newberry, Kimberly M. Newberry, Vu Ngo, Rosy Nguyen, Thai B. Nguyen, Tung Nguyen, Andrew Nishida, William S. Noble, Catherine S. Novak, Eva Maria Novoa, Briana Nuñez, Charles W. O'Donnell, Sara Olson, Kathrina C. Onate, Ericka Otterman, Hakan Ozadam, Michael Pagan, Tsultrim Palden, Xinghua Pan, Yongjin Park, E. Christopher Partridge, Benedict Paten, Florencia Pauli-Behn, Michael J. Pazin, Baikang Pei, Len A. Pennacchio, Alexander R. Perez, Emily H. Perry, Dmitri D. Pervouchine, Nishigandha N. Phalke, Quan Pham, Doug H. Phanstiel, Ingrid Plajzer-Frick, Gabriel A. Pratt, Henry E. Pratt, Sebastian Preissl, Jonathan K. Pritchard, Yuri Pritykin, Michael J. Purcaro, Qian Qin, Giovanni Quinones-Valdez, Ines Rabano, Ernest Radovani, Anil Raj, Nisha Rajagopal, Oren Ram, Lucia Ramirez, Ricardo N. Ramirez, Dylan Rausch, Soumya Raychaudhuri, Joseph Raymond, Rozita Razavi, Timothy E. Reddy, Thomas M. Reimonn, Bing Ren, Alexandre Reymond, Alex Reynolds, Suhn K. Rhie, John Rinn, Miguel Rivera, Juan Carlos Rivera-Mulia, Brian S. Roberts, Jose Manuel Rodriguez, Joel Rozowsky, Russell Ryan, Eric Rynes, Denis N. Salins, Richard Sandstrom, Takayo Sasaki, Shashank Sathe, Daniel Savic, Alexandra Scavelli, Jonathan Scheiman, Christoph Schlaffner, Jeffery A. Schloss, Frank W. Schmitges, Lei Hoon See, Anurag Sethi, Manu Setty, Anthony Shafer, Shuo Shan, Eilon Sharon, Quan Shen, Yin Shen, Richard I. Sherwood, Minyi Shi, Sunyoung Shin, Noam Shoresh, Kyle Siebenthall, Cristina Sisu, Teri Slifer, Cricket A. Sloan, Anna Smith, Valentina Snetkova, Michael P. Snyder,

Damek V. Spacek, Sharanya Srinivasan, Rohith Srivas, George Stamatoyannopoulos, John A. Stamatoyannopoulos, Rebecca Stanton, Dave Steffan, Sandra Stehling-Sun, J. Seth Strattan, Amanda Su, Balaji Sundararaman, Marie-Marthe Suner, Tahin Syed, Matt Szynkarek, Forrest Y. Tanaka, Danielle Tenen, Mingxiang Teng, Jeffrey A. Thomas, Dave Toffey, Michael L. Tress, Diane E. Trout, Gosia Trynka, Junko Tsuji, Sean A. Upchurch, Oana Ursu, Barbara Uszczynska-Ratajczak, Mia C. Uziel, Alfonso Valencia, Benjamin Van Biber, Arjan G. van der Velde, Eric L. Van Nostrand, Yekaterina Vaydylevich, Jesus Vazquez, Alec Victorsen, Jost Vielmetter, Jeff Vierstra, Axel Visel, Anna Vlasova, Christopher M. Vockley, Simona Volpi, Shinny Vong, Hao Wang, Mengchi Wang, Qin Wang, Ruth Wang, Tao Wang, Wei Wang, Xiaofeng Wang, Yanli Wang, Nathaniel K. Watson, Xintao Wei, Zhijie Wei, Hendrik Weisser, Sherman M. Weissman, Rene Welch, Robert E. Welikson, Zhiping Weng, Harm-Jan Westra, John W. Whitaker, Collin White, Kevin P. White, Andre Wildberg, Brian A. Williams, David Wine, Heather N. Witt, Barbara Wold, Maxim Wolf, James Wright, Rui Xiao, Xinshu Xiao, Jie Xu, Jinrui Xu, Koon-Kiu Yan, Yongqi Yan, Hongbo Yang, Xinqiong Yang, Yi-Wen Yang, Galip Gürkan Yardımcı, Brian A. Yee, Gene W. Yeo, Taylor Young, Tianxiong Yu, Feng Yue, Chris Zaleski, Chongzhi Zang, Haoyang Zeng, Weihua Zeng, Daniel R. Zerbino, Jie Zhai, Lijun Zhan, Ye Zhan, Bo Zhang, Jialing Zhang, Jing Zhang, Kai Zhang, Lijun Zhang, Peng Zhang, Qi Zhang, Xiao-Ou Zhang, Yanxiao Zhang, Zhizhuo Zhang, Yuan Zhao, Ye Zheng, Guoqing Zhong, Xiao-Qiao Zhou, Yun Zhu, Jared Zimmerman, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2493-4.

[114] Marina Lizio, Jayson Harshbarger, Hisashi Shimoji, Jessica Severin, Takeya Kasukawa, Serkan Sahin, Imad Abugessaisa, Shiro Fukuda, Fumi Hori, Sachi Ishikawa-Kato, Christopher J Mungall, Erik Arner, J Kenneth Baillie, Nicolas Bertin, Hidemasa Bono, Michiel de Hoon, Alexander D Diehl, Emmanuel Dimont, Tom C Freeman, Kaori Fujieda, Winston Hide, Rajaram Kaliyaperumal, Toshiaki Katayama, Timo Lassmann, Terrence F Meehan, Koro Nishikata, Hiromasa Ono, Michael Rehli, Albin Sandelin, Erik A Schultes, Peter A C 't Hoen, Zuotian Tatum, Mark Thompson, Tetsuro Toyoda, Derek W Wright, Carsten O Daub, Masayoshi Itoh, Piero Carninci, Yoshihide Hayashizaki, Alistair R R Forrest, Hideya Kawaji, and FANTOM consortium. Gate-

ways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*, 16(1):22, 2015. ISSN 1465-6906. doi: 10.1186/s13059-014-0560-6.

[115] Christina J Herrmann, Ralf Schmidt, Alexander Kanitz, Panu Artimo, Andreas J Gruber, and Mihaela Zavolan. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3 end sequencing. *Nucleic Acids Research*, 48(D1):D174–D179, 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz918.

[116] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M. Beckmann, Claudia Strein, Norman E. Davey, David T. Humphreys, Thomas Preiss, Lars M. Steinmetz, Jeroen Krijgsveld, and Matthias W. Hentze. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, 149(6):1393–1406, 2012. ISSN 0092-8674. doi: 10.1016/j.cell.2012.04.031.

[117] S B Martins, T Eide, R L Steen, T Jahnsen, S Skalhegg B, and P Collas. HA95 is a protein of the chromatin and nuclear matrix regulating nuclear envelope dynamics. *Journal of Cell Science*, 113(21):3703–3713, 2000. ISSN 0021-9533. doi: 10.1242/jcs.113.21.3703.

[118] Nigel G. Laing, Danielle E. Dye, Carina Wallgren-Pettersson, Gabriele Richard, Nicole Monnier, Suzanne Lillis, Thomas L. Winder, Hanns Lochmüller, Claudio Graziano, Stella Mitrani-Rosenbaum, Darren Twomey, John C. Sparrow, Alan H. Beggs, and Kristen J. Nowak. Mutations and polymorphisms of the skeletal muscle -actin gene (ACTA1). *Human Mutation*, 30(9):1267–1277, 2009. ISSN 1059-7794. doi: 10.1002/humu.21059.

[119] Sara S. Procknow and Beth A. Kozel. Emerging mechanisms of elastin transcriptional regulation. *American Journal of Physiology-Cell Physiology*, 323(3):C666–C677, 2022. ISSN 0363-6143. doi: 10.1152/ajpcell.00228.2022.

[120] Matthieu Lacroix, Geneviève Rodier, Olivier Kirsh, Thibault Houles, Hélène Delpech, Berfin Seyran, Laurie Gayte, Francois Casas, Laurence Pessemesse, Maud Heuillet, Floriant Bellvert, Jean-Charles Portais, Charlene Berthet, Florence Bernex, Michele Brivet, Audrey Boutron, Laurent Le Cam, and Claude Sardet. E4F1 controls a transcriptional program essential for pyruvate dehydrogenase activity. *Proceedings of the National Academy of Sciences*, 113(39):10998–11003, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1602754113.

[121] R A Kahn, F G Kern, J Clark, E P Gelmann, and C Rulka. Human ADP-ribosylation factors. A functionally conserved family of GTP-binding proteins. *Journal of Biological Chemistry*, 266(4):2606–2614, 1991. ISSN 0021-9258. doi: 10.1016/s0021-9258(18)52288-2.

[122] Michael L. Tress, Federico Abascal, and Alfonso Valencia. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, 42(2):98–110, 2017. ISSN 0968-0004. doi: 10.1016/j.tibs.2016.08.008.

[123] Julien Lagarde, Barbara Uszczynska-Ratajczak, Silvia Carbonell, Sílvia Pérez-Lluch, Amaya Abad, Carrie Davis, Thomas R Gingeras, Adam Frankish, Jennifer Harrow, Roderic Guigo, and Rory Johnson. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nature Genetics*, 49(12):1731–1740, 2017. ISSN 1061-4036. doi: 10.1038/ng.3988.

[124] Simon A. Hardwick, Wen Hu, Anoushka Joglekar, Li Fan, Paul G. Collier, Careen Foord, Jennifer Balacco, Samantha Lanjewar, Maureen McGuirk Sampson, Frank

Koopmans, Andrey D. Prjibelski, Alla Mikheenko, Natan Belchikov, Julien Jarroux, Anne Bergstrom Lucas, Miklós Palkovits, Wenjie Luo, Teresa A. Milner, Lishomwa C. Ndhlovu, August B. Smit, John Q. Trojanowski, Virginia M. Y. Lee, Olivier Fedrigo, Steven A. Sloan, Dóra Tombácz, M. Elizabeth Ross, Erich Jarvis, Zsolt Boldogkői, Li Gan, and Hagen U. Tilgner. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nature Biotechnology*, 40(7):1082–1092, 2022. ISSN 1087-0156. doi: 10.1038/s41587-022-01231-3.

[125] Feng Yue, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, Zhihai Ma, Carrie Davis, Benjamin D Pope, Yin Shen, Dmitri D Pervouchine, Sarah Djebali, Robert E Thurman, Rajinder Kaul, Eric Rynes, Anthony Kirilusha, Georgi K Marinov, Brian A Williams, Diane Trout, Henry Amrhein, Katherine Fisher-Aylor, Igor Antoshechkin, Gilberto DeSalvo, Lei-Hoon See, Meagan Fastuca, Jorg Drenkow, Chris Zaleski, Alex Dobin, Pablo Prieto, Julien Lagarde, Giovanni Bussotti, Andrea Tanzer, Olgert Denas, Kanwei Li, M A Bender, Miaohua Zhang, Rachel Byron, Mark T Groudine, David McCleary, Long Pham, Zhen Ye, Samantha Kuan, Lee Edsall, Yi-Chieh Wu, Matthew D Rasmussen, Mukul S Bansal, Manolis Kellis, Cheryl A Keller, Christapher S Morrissey, Tejaswini Mishra, Deepti Jain, Nergiz Dogan, Robert S Harris, Philip Cayting, Trupti Kawli, Alan P Boyle, Ghia Euskirchen, Anshul Kundaje, Shin Lin, Yiing Lin, Camden Jansen, Venkat S Malladi, Melissa S Cline, Drew T Erickson, Vanessa M Kirkup, Katrina Learned, Cricket A Sloan, Kate R Rosenbloom, Beatriz Lacerda de Sousa, Kathryn Beal, Miguel Pignatelli, Paul Flicek, Jin Lian, Tamer Kahveci, Dongwon Lee, W James Kent, Miguel Ramalho Santos, Javier Herrero, Cedric Notredame, Audra Johnson, Shinny Vong, Kristen Lee, Daniel Bates, Fidencio Neri, Morgan Diegel, Theresa Canfield, Peter J Sabo, Matthew S Wilken, Thomas A Reh, Erika Giste, Anthony Shafer, Tanya Kutyavin, Eric Haugen, Douglas Dunn, Alex P Reynolds, Shane Neph, Richard Humbert, R Scott Hansen, Marella De Bruijn, Licia Selleri, Alexander Rudensky, Steven Josefowicz, Robert Samstein, Evan E Eichler, Stuart H Orkin, Dana Levasseur, Thalia Papayannopoulou, Kai-Hsin Chang, Arthur Skoultchi, Srikanta Gosh, Christine Disteche, Piper Treuting, Yanli Wang, Mitchell J Weiss, Gerd A Blobel, Xiaoyi Cao, Sheng Zhong, Ting Wang, Peter J Good, Rebecca F Lowdon, Leslie B Adams, Xiao-Qiao Zhou, Michael J Pazin, Elise A Feingold, Barbara Wold, James Taylor, Ali Mortazavi, Sherman M Weissman, John A Stamatoyannopoulos, Michael P Snyder, Roderic Guigo, Thomas R Gingeras, David M Gilbert, Ross C Hardison, Michael A Beer, Bing Ren, and The Mouse ENCODE Consortium. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364, 2014. ISSN 0028-0836. doi: 10.1038/nature13992.

[126] Hazuki Takahashi, Sachi Kato, Mitsuyoshi Murata, and Piero Carninci. Gene Regulatory Networks, Methods and Protocols. *Methods in Molecular Biology*, 786:181–200, 2011. ISSN 1064-3745. doi: 10.1007/978-1-61779-292-2_11.

[127] Philippe Batut and Thomas R. Gingeras. RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5-Complete cDNAs. *Current Protocols in Molecular Biology*, 104(1):25B.11.1–25B.11.16, 2013. ISSN 1934-3639. doi: 10.1002/0471142727.mb25b11s104.

[128] Lingyun Song and Gregory E. Crawford. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian

Cells. *Cold Spring Harbor Protocols*, 2010(2):pdb.prot5384, 2010. ISSN 1940-3402. doi: 10.1101/pdb.prot5384.

[129] Juan L. Trincado, Juan C. Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J. Elliott, and Eduardo Eyras. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):40, 2018. ISSN 1474-7596. doi: 10.1186/s13059-018-1417-1.

[130] Dominik M. Endres and Johannes E. Schindelin. A New Metric for Probability Distributions. *IEEE Transactions on Information Theory*, 49(7):1858, 2003. ISSN 0018-9448. doi: 10.1109/tit.2003.813506.

[131] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1 0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A Nicholson, David R Hagen, Dmitrii V Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G Young, Gavin A Price, Gert-Ludwig Ingold, Gregory E Allen, Gregory R Lee, Hervé Audren, Irvin Probst, Jörg P Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A Brodtkorb, Perry Lee, Robert T McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J Pingel, Thomas P Robitaille, Thomas Spura, Thouis R Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. ISSN 1548-7091. doi: 10.1038/s41592-019-0686-2.

[132] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics*, 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac757.

[133] Zhuorui Xie, Allison Bailey, Maxim V. Kuleshov, Daniel J. B. Clarke, John E. Evangelista, Sherry L. Jenkins, Alexander Lachmann, Megan L. Wojciechowicz, Eryk Kropiwnicki, Kathleen M. Jagodnik, Minji Jeon, and Avi Ma'ayan. Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, 1(3):e90, 2021. ISSN 2691-1299. doi: 10.1002/cpz1.90.

[134] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):

403–410, 1990. ISSN 0022-2836. doi: 10.1016/s0022-2836(05)80360-2.

[135] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq033.

[136] Victoria Nembaware, Kenneth H. Wolfe, Fabiana Bettoni, Janet Kelso, and Cathal Seoighe. Allele-specific transcript isoforms in human. *FEBS Letters*, 577(1-2):233–238, 2004. ISSN 0014-5793. doi: 10.1016/j.febslet.2004.10.018.

[137] Zhi-Xiang Lu, Peng Jiang, and Yi Xing. Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdisciplinary Reviews: RNA*, 3(4):581–592, 2012. ISSN 1757-7004. doi: 10.1002/wrna.120.

[138] Robert K. Bradley and Olga Anczuków. RNA splicing dysregulation and the hallmarks of cancer. *Nature Reviews Cancer*, 23(3):135–155, 2023. ISSN 1474-175X. doi: 10.1038/s41568-022-00541-7.

[139] IGVF Consortium. The Impact of Genomic Variation on Function (IGVF) Consortium. *arXiv*, 2023. doi: 10.48550/arxiv.2307.13708.

[140] Gene Yeo, Dirk Holste, Gabriel Kreiman, and Christopher B Burge. Variation in alternative splicing across human tissues. *Genome Biology*, 5(10):R74, 2004. ISSN 1465-6906. doi: 10.1186/gb-2004-5-10-r74.

[141] Nuno L. Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y. Xiong, Serge Gueroussov, Leo J. Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Çolak, Tae-Hyung Kim, Christine M. Misquitta-Ali, Michael D. Wilson, Philip M. Kim, Duncan T. Odom, Brendan J. Frey, and Benjamin J. Blencowe. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, 338(6114):1587–1593, 2012. ISSN 0036-8075. doi: 10.1126/science.1230612.

[142] Manuel Irimia, Robert J. Weatheritt, Jonathan D. Ellis, Neelroop N. Parikshak, Thomas Gonatopoulos-Pournatzis, Mariana Babor, Mathieu Quesnel-Vallières, Javier Tapial, Bushra Raj, Dave O'Hanlon, Miriam Barrios-Rodiles, Michael J.E. Sternberg, Sabine P. Cordes, Frederick P. Roth, Jeffrey L. Wrana, Daniel H. Geschwind, and Benjamin J. Blencowe. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell*, 159(7):1511–1523, 2014. ISSN 0092-8674. doi: 10.1016/j.cell.2014.11.035.

[143] Bushra Raj and Benjamin J. Blencowe. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron*, 87(1):14–27, 2015. ISSN 0896-6273. doi: 10.1016/j.neuron.2015.05.004.

[144] Bongmin Bae and Pedro Miura. Emerging Roles for 3 UTRs in Neurons. *International Journal of Molecular Sciences*, 21(10):3413, 2020. doi: 10.3390/ijms21103413.

[145] Muhammad Irfan, Katisha R. Gopaul, Omid Miry, Tomas Hökfelt, Patric K. Stanton, and Christina Bark. SNAP-25 isoforms differentially regulate synaptic transmission and long-term synaptic plasticity at central synapses. *Scientific Reports*, 9(1):6403, 2019. doi: 10.1038/s41598-019-42833-3.

[146] Charlotte Soneson, Katarina L. Matthes, Malgorzata Nowicka, Charity W. Law, and Mark D. Robinson. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, 17(1):12, 2016. ISSN 1474-7596. doi: 10.1186/s13059-015-0862-3.

[147] Man Hagiyama, Naoki Ichiyanagi, Keiko B. Kimura, Yoshinori Murakami, and Akihiko

Ito. Expression of a Soluble Isoform of Cell Adhesion Molecule 1 in the Brain and Its Involvement in Directional Neurite Outgrowth. *The American Journal of Pathology*, 174(6):2278–2289, 2009. ISSN 0002-9440. doi: 10.2353/ajpath.2009.080743.

[148] Kasper Karlsson, Peter Lönnerberg, and Sten Linnarsson. Alternative TSSs are coregulated in single cells in the mouse brain. *Molecular Systems Biology*, 13(5):930, 2017. ISSN 1744-4292. doi: 10.15252/msb.20167374.

[149] Matthew S. Hestand, Andreas Klingenhoff, Matthias Scherf, Yavuz Ariyurek, Yolande Ramos, Wilbert van Workum, Makoto Suzuki, Thomas Werner, Gert-Jan B. van Ommen, Johan T. den Dunnen, Matthias Harbers, and Peter A.C. 't Hoen. Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. *Nucleic Acids Research*, 38(16):e165–e165, 2010. ISSN 0305-1048. doi: 10.1093/nar/gkq602.

[150] Bongmin Bae, Hannah N. Gruner, Maebh Lynch, Ting Feng, Kevin So, Daniel Oliver, Grant S Mastick, Wei Yan, Simon Pieraut, and Pedro Miura. Elimination of Calm1 long 3' UTR mRNA isoform by CRISPR-Cas9 gene editing impairs dorsal root ganglion development and hippocampal neuron activation in mice. *RNA*, 26(10):rna.076430.120, 2020. ISSN 1355-8382. doi: 10.1261/rna.076430.120.

[151] Anthony T. Campagnoni and Robert P. Skoff. The Pathobiology of Myelin Mutants Reveal Novel Biological Functions of the MBP and PLP Genes. *Brain Pathology*, 11 (1):74–91, 2001. ISSN 1015-6305. doi: 10.1111/j.1750-3639.2001.tb00383.x.

[152] Kazuhiro Ikenaka, Tetsushi Kagawa, and Katsuhiko Mikoshiba. Selective Expression of DM-20, an Alternatively Spliced Myelin Proteolipid Protein Gene Product, in Developing Nervous System and in Nonglial Cells. *Journal of Neurochemistry*, 58(6): 2248–2253, 1992. ISSN 0022-3042. doi: 10.1111/j.1471-4159.1992.tb10970.x.

[153] Yu Yamazaki, Na Zhao, Thomas R. Caulfield, Chia-Chen Liu, and Guojun Bu. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nature Reviews Neurology*, 15(9):501–518, 2019. ISSN 1759-4758. doi: 10.1038/s41582-019-0228-7.

[154] Y Yamamoto, R Mizuno, T Nishimura, Y Ogawa, H Yoshikawa, H Fujimura, E Adachi, T Kishimoto, T Yanagihara, and S Sakoda. Cloning and expression of myelin-associated oligodendrocytic basic protein. A novel basic protein constituting the central nervous system myelin. *Journal of Biological Chemistry*, 269(50):31725–31730, 1994. ISSN 0021-9258. doi: 10.1016/s0021-9258(18)31756-3.

[155] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *bioRxiv*, page 2022.12.14.520412, 2022. doi: 10.1101/2022.12.14.520412.

[156] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, Kerstin Howe, Miten Jain, Tsung-Yu Lu, Charles Markello, Fergal J. Martin, Matthew W. Mitchell, Katherine M. Munson, Moses Njagi Mwaniki, Adam M. Novak, Hugh E. Olsen, Trevor Pesout, David Porubsky, Pjotr Prins, Jonas A. Sibbesen, Jouni Sirén, Chad Tomlinson, Flavia Villani, Mitchell R. Vollger, Lucinda L. Antonacci-Fulton, Gunjan Baid, Carl A. Baker, Anas-

tasiya Belyaeva, Konstantinos Billis, Andrew Carroll, Pi-Chuan Chang, Sarah Cody, Daniel E. Cook, Robert M. Cook-Deegan, Omar E. Cornejo, Mark Diekhans, Peter Ebert, Susan Fairley, Olivier Fedrigo, Adam L. Felsenfeld, Giulio Formenti, Adam Frankish, Yan Gao, Nanibaa' A. Garrison, Carlos Garcia Giron, Richard E. Green, Leanne Haggerty, Kendra Hoekzema, Thibaut Hourlier, Hanlee P. Ji, Eimear E. Kenny, Barbara A. Koenig, Alexey Kolesnikov, Jan O. Korbel, Jennifer Kordosky, Sergey Koren, HoJoon Lee, Alexandra P. Lewis, Hugo Magalhães, Santiago Marco-Sola, Pierre Marijon, Ann McCartney, Jennifer McDaniel, Jacquelyn Mountcastle, Maria Nattestad, Sergey Nurk, Nathan D. Olson, Alice B. Popejoy, Daniela Puiu, Mikko Rautiainen, Allison A. Regier, Arang Rhie, Samuel Sacco, Ashley D. Sanders, Valerie A. Schneider, Baergen I. Schultz, Kishwar Shafin, Michael W. Smith, Heidi J. Sofia, Ahmad N. Abou Tayoun, Françoise Thibaud-Nissen, Francesca Floriana Tricomi, Justin Wagner, Brian Walenz, Jonathan M. D. Wood, Aleksey V. Zimin, Guillaume Bourque, Mark J. P. Chaisson, Paul Flicek, Adam M. Phillippy, Justin M. Zook, Evan E. Eichler, David Haussler, Ting Wang, Erich D. Jarvis, Karen H. Miga, Erik Garrison, Tobias Marschall, Ira M. Hall, Heng Li, and Benedict Paten. A draft human pangenome reference. *Nature*, 617(7960):312–324, 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-05896-x.

[157] Jordan M. Eizenga, Adam M. Novak, Jonas A. Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, Xian Chang, Josiah D. Seaman, Robin Rounthwaite, Jana Ebler, Mikko Rautiainen, Shilpa Garg, Benedict Paten, Tobias Marschall, Jouni Sirén, and Erik Garrison. Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, 21 (1):139–162, 2020. ISSN 1527-8204. doi: 10.1146/annurev-genom-120219-080406.

[158] Renee Salz, Nuno Saraiva-Agostinho, Emil Vorsteveld, Caspar I. van der Made, Simone Kersten, Merel Stemerdink, Jamie Allen, Pieter-Jan Volders, Sarah E. Hunt, Alexander Hoischen, and Peter A.C. 't Hoen. SUsPECT: a pipeline for variant effect prediction based on custom long-read transcriptomes for improved clinical variant annotation. *BMC Genomics*, 24(1):305, 2023. doi: 10.1186/s12864-023-09391-5.

[159] Jonas A. Sibbesen, Jordan M. Eizenga, Adam M. Novak, Jouni Sirén, Xian Chang, Erik Garrison, and Benedict Paten. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nature Methods*, 20(2):239–247, 2023. ISSN 1548-7091. doi: 10.1038/s41592-022-01731-9.

[160] Wend Yam Donald Davy Ouedraogo and Aida Ouangraoua. Comparative Genomics, 20th International Conference, RECOMB-CG 2023, Istanbul, Turkey, April 14–15, 2023, Proceedings. *Lecture Notes in Computer Science*, pages 19–34, 2023. ISSN 0302-9743. doi: 10.1007/978-3-031-36911-7_2.

[161] Safa Jammali, Abigaïl Djossou, Wend-Yam D D Ouédraogo, Yannis Nevers, Ibrahim Chegrane, and Aïda Ouangraoua. From pairwise to multiple spliced alignment. *Bioinformatics Advances*, 2(1):vbab044, 2022. doi: 10.1093/bioadv/vbab044.

[162] Karan Bedi, Brian Magnuson, Ishwarya Venkata Narayanan, Michelle T. Paulsen, Thomas E. Wilson, and Mats Ljungman. Cotranscriptional splicing efficiencies differ within genes and between cell types. *RNA*, 27(7):829–840, 2021. ISSN 1355-8382. doi: 10.1261/rna.078662.120.

[163] Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E. Kleinman,

Thomas M. Hyde, Nikhil Rao, Stephanie C. Hicks, Keri Martinowich, and Andrew E. Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, 2021. ISSN 1097-6256. doi: 10.1038/s41593-020-00787-0.

[164] Emily Miyoshi, Samuel Morabito, Caden M Henningfield, Negin Rahimzadeh, Sepideh Kiani Shabestari, Sudeshna Das, Neethu Michael, Fairlie Reese, Zechuan Shi, Zhenkun Cao, Vanessa Scarfone, Miguel A Arreola, Jackie Lu, Sierra Wright, Justine Silva, Kelsey Leavy, Ira T Lott, Eric Doran, William H Yong, Saba Shahin, Mari Perez-Rosendahl, Elizabeth Head, Kim N Green, and Vivek Swarup. Spatial and single-nucleus transcriptomic analysis of genetic and sporadic forms of Alzheimer's Disease. 2023. doi: 10.1101/2023.07.24.550282.

[165] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp616.

[166] Matthias Schaefer, Utkarsh Kapoor, and Michael F. Jantsch. Understanding RNA modifications: the promises and technological bottlenecks of the 'epitranscriptome'. *Open Biology*, 7(5):170077, 2017. doi: 10.1098/rsob.170077.

[167] Adrien Leger, Paulo P. Amaral, Luca Pandolfini, Charlotte Capitanchik, Federica Capraro, Valentina Miano, Valentina Migliori, Patrick Toolan-Kerr, Theodora Sideri, Anton J. Enright, Konstantinos Tzelepis, Folkert J. van Werven, Nicholas M. Luscombe, Isaia Barbieri, Jernej Ule, Tomas Fitzgerald, Ewan Birney, Tommaso Leonardi, and Tony Kouzarides. RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nature Communications*, 12(1):7198, 2021. doi: 10.1038/s41467-021-27393-3.

[168] Veronica Dezi, Chavdar Ivanov, Irmgard U Haussmann, and Matthias Soller. Nucleotide modifications in messenger RNA and their role in development and disease. *Biochemical Society Transactions*, 44(5):1385–1393, 2016. ISSN 0300-5127. doi: 10.1042/bst20160110.

[169] Miten Jain, Robin Abu-Shumays, Hugh E. Olsen, and Mark Akeson. Advances in nanopore direct RNA sequencing. *Nature Methods*, 19(10):1160–1164, 2022. ISSN 1548-7091. doi: 10.1038/s41592-022-01633-w.

[170] Oxford Nanopore. Oxford nanopore london calling 2023 technology update, 2023. URL `https://nanoporetech.com/oxford-nanopore-london-calling-23-technology-update#direct-rna`.