UNIVERSITY OF CALIFORNIA,
IRVINE

Full characterization of transcriptomes using long read sequencing

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational, and Systems Biology

by

Dana Elizabeth Wyman

Dissertation Committee:
Professor Ali Mortazavi, Chair
Assistant Professor Zeba Wunderlich
Associate Professor Robert Spitale
Assistant Professor J.J. Emerson
Professor Klemens Hertel

2020

DEDICATION

To

Ossian O'Reilly and the Wyman family

in recognition of their love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to offer my deepest thanks to my PhD advisor and chair, Dr. Ali Mortazavi, for encouraging my interest in research as an undergraduate summer student all the way through my PhD. His support and mentorship have been instrumental in helping me develop into the scientist I am today.

I would like to thank my committee members, Dr. Zeba Wunderlich, Dr. Robert Spitale. Dr. J.J. Emerson, and Dr. Klemens Hertel. Their guidance and thoughtful questions have contributed greatly to my work.

I owe a great debt of gratitude to the ENCODE consortium, which has not only funded my work but has also provided opportunities for growth and collaboration within its talented community. Specifically, I would like to thank Dr. Barbara Wold and her laboratory, as well as the ENCODE Data Coordination Center team.

I thank the University of California Irvine Center for Complex Biological Systems, including the Mathematical, Computational, and Systems Biology program. Karen Martin deserves special recognition for going above and beyond to ensure the well-being of graduate students like me.

I would like to thank my husband, Ossian O'Reilly, for sharing both the highs and lows of my PhD with me. I would also like to thank my parents, my sister, Sarah, and my extended family for their lifelong support and encouragement. You have helped me achieve my academic goals while also reminding me to keep perspective during challenging times.

Last but not least, I would like to thank my friends both inside and outside of the lab for enriching my life immeasurably. Special thanks to Natalie Telis for being an amazing friend and career role model. Thank you to Gaby Balderrama-Gutierrez and Fairlie Reese for their friendship and admirable commitment to the TALON project. Thank you to Kate Williams for her encouragement and humor. Thank you to every member of the Mortazavi lab for your kindness, support, and for making the lab such an exceptional place to work.

VITA

Dana Elizabeth Wyman

2020 PhD in Mathematical, Computational, and Systems Biology

2016 M.S. in Biomedical Informatics, Stanford University

2014 B.S. with Honors in Biology. Focus area: Molecular and cell biology

FIELD OF STUDY

Mathematical, Computation, and Systems Biology

PUBLICATIONS

Wyman, D.*, Williams, B.A *, Balderrama-Gutierrez, G., McGill, C. M., Trout, D., Wold, B.J.*, & Mortazavi, A.* (2020). A single-cell, isoform-level survey of the developing mouse forelimb using deep long-read sequencing. In preparation.

Wyman, D.*, Balderrama-Gutierrez, G.*, Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B.A., Trout, D., England, W., Chu, S., Spitale, R.C., Tenner, A.J., Wold, B.J., & Mortazavi, A. (2020). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv 672931.

Wyman, D., & Mortazavi, A. (2019). TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, *35*(2), 340–342.

Rahmanian, S., Balderrama-Gutierrez, G., **Wyman, D**., McGill, C. M., Nguyen, K., Spitale, R.C., & Mortazavi, A. (2020). Long-TUC-seq: A robust method for quantification of metabolically labeled full-length isoforms. In preparation.

Spitale, R.C., Chan, D., Feng, C., England, W., **Wyman, D.**, Flynn, R., Wang, X., Shi, Y. & Mortazavi, A. (2020). Transcriptome-Wide Combinatorial RNA Structure Probing in Living Cells. bioRxiv 2020.03.24.006866.

Ramirez, R. N., El-Ali, N. C., Anne Mager, M., **Wyman, D.**, Conesa, A., & Mortazavi, A. (2017). Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. *Cell Systems*, *4*, 416–429.e3.

ABSTRACT OF THE DISSERTATION

Full characterization of transcriptomes using long read sequencing

By

Dana Elizabeth Wyman

Doctor of Philosophy in Mathematical, Computational, and Systems Biology

University of California, Irvine, 2020

Professor Ali Mortazavi, Chair

Almost all multi-exonic human genes are believed to undergo alternative splicing, giving rise to isoforms with potentially distinct functions, tissue specificities, and developmental roles. Differential isoform usage has been implicated in both normal developmental processes and in disease states. Much of the previous work attempting to identify and to quantify individual gene isoforms has been performed using short-read RNA sequencing on the Illumina platform. While this technology is considered the state of the art for quantifying gene expression, short reads are unable to accurately resolve full-length mammalian isoforms, which can be multiple kilobases long. Although computational methods have been developed to reconstruct isoforms from short reads, these are not able to overcome the fundamental limitations of the technology.

Long-read sequencing platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) bypass the transcript reconstruction challenges of short reads and offer the additional advantage of sequencing single molecules individually. PacBio sequencing in particular has been used extensively for de novo isoform reconstruction but was previously not deemed useful for quantitative measurements of gene or transcript expression due

both to the cost of the assay and to its relatively low throughput. However, technical advances have increased the yield as well as the accuracy of longer reads, presenting an opportunity to use these technologies directly for isoform-level quantification. Here, I present novel methods for long-read error correction, isoform discovery, and quantification in RNA samples from both pooled and single cells. First, I introduce TranscriptClean, a program that leverages a reference genome to correct common sequencing errors in long reads. Next, I describe TALON, a technology-agnostic approach to discovering and quantifying isoforms in multiple long-read datasets. Finally, I apply TALON to the analysis of deeply sequenced single cells from the developing mouse limb bud, demonstrating that long reads can provide key biological insights in the context of development. Together, these projects help pave the way for long-read transcriptome analyses on both the bulk and single-cell level, which grant us new insights into isoform expression across diverse human and mouse tissues.

Chapter 1

Introduction: Characterizing gene and isoform expression using third-generation sequencing technologies

Chapter 1

Introduction: Characterizing gene and isoform expression using third-generation sequencing technologies

1.1. Abstract

The process of alternative splicing can generate isoforms from the same gene with potentially different biological functions. Depending on the context in which an isoform is expressed, it may contribute to normal biological processes or to a disease state. Unfortunately, the most common method for profiling gene expression, short-read RNA-seq, cannot readily distinguish isoforms due to fundamental limits of the technology. However, long-read sequencing platforms such as PacBio and Oxford Nanopore have the potential to greatly improve our understanding of isoform-level RNA expression in the context of both normal development and disease. Here, I review different methods for studying alternative splicing, with particular focus on the current use of long-read sequencing in the field. Additionally, I describe how single-cell RNA sequencing (scRNA-seq) methods have been adapted to work with long-read technologies.

1.2. Introduction

1.2.1. Alternative splicing in development and disease

Differences in gene expression are essential to shaping the wide variety of cell phenotypes present in an organism throughout its lifetime[1]. In eukaryotes, the composition of the transcriptome is further modified by the process of alternative splicing. During splicing, intronic sequences and certain exons are excised from the mRNA transcript, thus expanding the number of possible transcripts that a single gene can code for[2]. In the human

genome, approximately ~20,000 protein coding genes are believed to produce more than 100,000 isoforms through alternative splicing, greatly expanding the potential diversity of the proteome[3]. Depending on the exons they contain, isoforms from the same gene may have similar, distinct, or even antagonistic functions[4,5]. In one striking case, the *BCL-X$_S$* isoform of the *BCL2L1* gene is known to promote apoptosis, while a second alternatively spliced isoform, BCL-X$_L$, instead inhibits apoptosis and is expressed in many cancers[6,7]. Some isoforms have been observed to be very specific to a particular tissue or developmental stage[8–10]. This type of specificity may be achieved through the context-specific expression of different RNA-binding proteins (RBPs), which bind to sequence motifs within mRNA molecules and help direct the splicing machinery[11]. For example, the *RBFOX* and *NOVA* families of splicing factors promote inclusion of specific exons in developing human and mouse neurons[12,13]. *NOVA* in particular is known to regulate splicing of genes involved in maintaining synaptic plasticity[14]. Aberrant isoform expression patterns have been implicated in multiple diseases, particularly in tissues such as the brain where alternative splicing is particularly prolific[12,15]. For instance, the relative proportions of *Mapt* isoforms in brains affected by Alzheimer's disease deviate from normal levels, favoring a class of transcripts that contain an extra microtubule binding domain[16].

1.2.2. Application of RNA-seq to the study of isoforms and alternative splicing

Over the years, many methods have been applied to the twin problems of characterizing and quantifying alternative splicing and exon usage. The first of these, the northern blot, was introduced in 1977. This technique detects mRNA from a particular gene or isoform by first isolating the RNA by size, affixing it to a membrane, and then

visualizing the signal by hybridizing the RNA to labeled probes[17,18]. Beginning in the 1990s, the reverse transcription polymerase chain reaction (RT-PCR) technique was applied to quantify isoforms[19–21]. RT-PCR is still considered a gold standard for validating gene expression measurements today. On the characterization side, Sanger sequencing of expressed sequence tags (ESTs) made it possible to identify new splicing events and call novel genes[22]. However, northern blots, RT-PCR, and EST sequencing are considered low-throughput techniques and are therefore not readily applicable to full transcriptomes[23]. Isoform-specific microarrays address this scalability problem by arranging thousands of transcript or exon-specific probes on a chip, allowing massively parallel measurements to be made on a transcriptome-wide level[24]. However, microarrays suffer from challenges of their own. The signal tends to be noisy, and it is often challenging to design probes to accurately detect specific isoforms of a gene[23]. And although it is possible to design probes to screen for novel splicing provided that the flanking exons are known, microarrays are mostly limited to detecting known events.

The development of high-throughput, short-read RNA sequencing methods (RNA-seq) around 2008 completely changed the way gene and isoform expression are studied today. Although individual workflows vary, RNA-seq broadly involves reverse-transcribing RNA to cDNA, and fragmenting the cDNA into short, uniform pieces (50-300bp), followed by amplification and high-throughput sequencing[25,26]. The resulting reads can be aligned to the reference transcriptome or genome and counted in order to provide a quantitative expression level for both known and unknown genes[25,26]. Cheap and relatively accurate, RNA-seq has been widely adopted across biological and biomedical disciplines as the

predominant way to study gene expression. This has included large consortia such as the Encyclopedia of DNA elements (ENCODE). As part of its mission to characterize the functions of all regions in the genome, ENCODE has generated and released hundreds of RNA-seq datasets from a variety of human and mouse cell types[27]. While ENCODE largely focuses on phenotypically normal samples, RNA-seq has also proved a powerful tool for characterizing gene expression in disease. For instance, the Cancer Genome Atlas (TCGA) has amassed RNA-seq data from thousands of patient samples in over 30 cancer types[28]. This is an invaluable resource for the cancer research community, and has helped advance breakthroughs in the diagnosis, prevention, and treatment of cancer[29,30].

Short-read RNA-seq data has also been widely used to study alternative splicing[31]. For instance, computational methods have been developed to quantify differential exon usage across samples from RNA-seq data[32]. Known and novel splice junctions can also be quantified from RNA-seq data by counting the number of reads spanning the junctions in question[33–35]. However, recovering alternatively spliced junctions of a full isoform is challenging with this technology. The core problem is that short-read sequencing requires cDNA transcripts to be cut into 50-300bp pieces and amplified many times over **(Figure 1.1)**. Since mammalian transcripts routinely measure multiple kilobases in length, short reads seldom span more than one splice junction at a time[8]. This makes it very difficult to tell which exons were originally present in the source transcript. A series of bioinformatic methods have been devised to reconstruct and quantify isoforms from short-read data. For example, the MAJIQ method developed by Vaquero-Garcia *et al*. assembles a splicing graph from the transcriptome annotation, then models alternative isoform usage across samples

and conditions by calling local splice variation events for each gene from short reads[36]. The rMATS package provides a Bayesian statistical framework for modeling isoform uncertainty in RNA-seq measurements and incorporates biological replicates to help increase statistical power to call differential isoform events [37]. However, one major drawback of rMATS is that it uses a two-isoform model for each gene, which does not realistically capture the true complexity of alternative splicing. The Kallisto program takes a different approach by constructing a de Bruijn graph from k-mers of reference transcript sequences, and then pseudo-aligning short reads to these graphs[38]. Expectation maximization is used to assign reads to the isoforms that they are most likely to originate from[38]. While this algorithm is able to decide which isoforms are consistent with the observed short-read data, it cannot resolve the ground truth of exactly which isoforms were originally present. In addition, this method is limited by its dependence on the choice of transcriptome reference and can only measure known isoforms. Overall, RNA-seq has proved useful for identifying alternative exon and splice junction usage but falls short with respect to identifying complete splicing patterns in full-length transcripts.

1.2.3. Emerging long-read sequencing technologies

For over a decade, next-generation, short-read sequencing has reigned supreme in genomics. However, it is far from the only option available. Since as early as 2010, third-generation sequencing platforms such as PacBio and Oxford Nanopore have pioneered the use of long reads in the field[39,40]. These platforms are in principle ideal for sequencing isoforms because they produce reads long enough to capture entire mRNA transcripts at single-molecule resolution. PacBio has a maximum read length of 60 kb, while Oxford

Nanopore has been known to generate reads up to 1 Mb[41]. However, both technologies have historically been plagued by two major problems: low throughput in terms of read count and indel-driven error rates of up to 15-20%[39]. Compared to a single-pass error rate of less than 1% in Illumina short-read sequencing, this is exceedingly high[41]. The sources of these errors differ in PacBio and Oxford Nanopore due to underlying differences in the technology. In a PacBio run, cDNA molecules are first diffused into a series of individual zero-mode wave guide (ZMW) structures containing a DNA polymerase. As DNA synthesis proceeds, fluorescently labeled nucleotides are incorporated into the growing sequence in each ZMW, and the emitted pulses of light are captured on video[39]. Nucleotides are identified in real time based on the emission profile of the light pulse generated while the base resides with the polymerase[39]. Indels and mismatches are introduced into the sequence when pulses are missed, duplicated, or when the emission profile is mistakenly attributed to the wrong base[39]. These errors are considered random rather than related to the underlying composition of the source sequence, which means that additional read coverage can help improve accuracy[39,41,42].

Oxford Nanopore, on the other hand, has long struggled with non-random, sequence-specific bias in their errors[43]. In this technology, individual DNA or RNA molecules are pulled through bioengineered protein pores spanning a membrane with an electrical gradient[40]. As nucleotides pass through the pore, the change in electrical current is measured and bases are called from the characteristic changes in signal[40]. Approximately five connected nucleotides fit inside a pore at any given time, so basecalling must be performed on this sliding window rather than on individual bases[43,44]. Early Nanopore

basecallers used the Viterbi algorithm or hidden Markov models for this purpose, while newer versions apply deep neural networks[44–46]. Some sequence combinations, such as homopolymers, are more difficult to call correctly than others, which means that Nanopore errors come from systematic as well as random sources[47]. Nevertheless, Oxford Nanopore's approach to sequencing offers some substantial advantages over PacBio. Since it does not rely on sequencing by synthesis, Oxford Nanopore is able to sequence native RNA molecules directly without amplification. Direct-RNA sequencing avoids the PCR artifacts and amplification biases seen in cDNA technologies and offers the added benefit of detecting RNA modifications[48]. In addition, Oxford Nanopore has a much lower capital cost and smaller device footprint compared to PacBio. This has sped its adoption by research groups around the world and allowed sequencing to be performed in real time in environments ranging from rainforests to the International Space Station[49,50].

Significant efforts have been made to reduce long-read error rates. The most fundamental of these approaches, consensus sequencing, works by comparing multiple sequencing passes over the same source molecule in order to more accurately call the read sequence. In the case of PacBio, this is achieved by ligating circular adaptors onto the blunt ends of each double-stranded cDNA during library preparation[51]. This template opens up into a circle during sequencing, allowing the DNA polymerase to continuously sequence the insert molecule **(Figure 1.2)**. After sequencing, the PacBio circular consensus (CCS) software splits the read into constituent subreads based on the delimiting adaptor sequences, and compares them in order to arrive at a single consensus sequence for the insert[52]. Since PacBio has a stochastic error profile, CCS is highly effective in correcting

single-pass errors, but it should be noted that its success depends on obtaining sufficient numbers of sequencing passes[41]. CCS therefore becomes less effective as the insert molecule increases in length. Consensus sequencing methods have also been developed for Oxford Nanopore. For instance, the R2C2 method from Volden *et al*. uses a molecular cloning method called Gibson assembly to circularize the cDNA transcript, and amplifies it to create multiple consecutive copies of the insert for Nanopore sequencing[53]. However, the non-random errors generated by Oxford Nanopore are a continuing obstacle, since consensus correction is less likely to succeed when there is an underlying sequence bias. Also, this method only works on cDNA, so it is not useful for mitigating direct-RNA sequencing errors.

In spite of the associated challenges, third-generation sequencing technologies have been widely used to study alternative splicing. Given the high cost and low throughputs of the original platforms, many studies have used PacBio to assemble a catalogue of full-length isoforms in the sample of interest, and then mapped matching short-read data to this reference transcriptome in order to quantify isoform expression[52,54–56]. The PacBio-affiliated transcriptome pipeline, Isoseq, is designed for this type of analysis. Isoseq performs *de novo* isoform assembly by clustering PacBio reads in a sample and collapsing them to form distinct transcript models[52]. One advantage of this approach is that it can be applied in organisms that lack a high-quality reference genome, making it possible to identify genes in less well-studied species. However, the Isoseq paradigm is less ideal in settings where reference availability is not a problem, such as for human and mouse. For example, clustering-based methods have been known to collapse transcripts from highly

similar gene families into the same model, and they routinely merge shorter isoforms from the same gene into longer ones[57]. More concerningly, isoform polishing steps obscure interesting sequence differences such as small variants and RNA editing events that may have occurred in different reads. Finally, different clustering runs on the same data may generate substantially different results depending on how the clusters were initialized and merged, leading to problems with reproducibility[58].

Recently, technical advances have increased the yield as well as the accuracy of long reads, opening up new avenues for isoform study. To give a sense of the scale involved, the newest PacBio Sequel 2 machine produces up to 8 million reads per sequencing unit compared to 150,000 on the older RSII machines[59]. Comparable improvements have been made on the Oxford Nanopore side as well[59]. This progress towards more cost-effective, deep long-read sequencing raises the possibility of directly quantifying isoform-level expression from PacBio and Oxford Nanopore reads themselves. However, computational challenges remain. Most long-read software packages were not originally designed for this purpose, relying on read clustering or assembly approaches for calling isoforms. For example, the SQANTI package was originally developed to perform quality control on transcript models assembled by the PacBio Isoseq pipeline, and uses isoform abundance estimates from the Isoseq collapsing step as input for quantification[60]. A different program, FLAIR, uses the Minimap2 aligner to map long reads to each other in order to create a consolidated isoform model set, then assigns the reads to these models after various sequencing error correction steps[61]. The StringTie2 package focuses on the problem of assembling full-length isoforms from multiple long reads much the way one would with

short reads, but this approach has the disadvantage of chaining together reads that might represent distinct isoforms in their own right rather than originating from one molecule[62]. Another problem is that many long-read software packages were developed with either PacBio or Oxford Nanopore in mind, even if they can technically be made to run on both. In light of this, new methods are needed to process long read data from different platforms in a way that that allows for simultaneous transcript discovery and quantification in both regular bulk RNA samples and single cells.

1.2.4. Single-cell transcriptomics

Thus far, our discussion has centered on bulk, pooled-cell approaches to characterizing gene and isoform expression. However, in some systems, bulk sequencing is not sufficient to understand the biology at hand. When thousands or millions of cells are processed together, the resulting expression measurements are effectively averaged across the sample, obscuring interesting cell-cell differences that may have significant biological implications[63]. This is especially problematic when the sample in question consists of heterogenous cell types or when the cell type of interest is rare in the population[64]. For instance, patient tumor samples commonly used for cancer research typically contain a mix of cell types, ranging from the cancerous tissue itself to blood vessels, immune cells, and phenotypically normal surrounding tissue. Considerable heterogeneity may also exist among cells that nominally belong to the same type. For example, T-cell receptor diversity is essential in order to detect and destroy invading pathogens[65]. Bulk RNA-seq experiments conducted on such samples are likely to average out interesting differences between cells and subpopulations.

In light of these challenges, single-cell RNA sequencing approaches have been developed to allow transcriptome analysis to take place on the level of individual cells[66–69]. These can be broadly divided into two classes: those that amplify transcripts in full (i.e. SMART-seq, SPLiT-seq)[67,70], and methods that tag and sequence only the 3' or the 5' end of transcripts (i.e. Drop-Seq, 10X Chromium, STRT-seq)[68,71,72]. These methods come with important tradeoffs. Full-length, single-cell preparations preserve transcript information, but are historically low-throughput with respect to cell count because they require cells to be placed in individual wells. The combinatorial barcoding approach introduced by SPLiT-seq represents a potential improvement here, since it does not require physical separation of the cells[70]. Conversely, end-tagging approaches allow tens of thousands of cells to be sequenced at a time, and have been leveraged by consortia such as the Human Cell Atlas[73] and Tabula Muris[74] to exhaustively catalogue cell subtypes in human and mouse. However, 3'-end methods necessarily lack any kind of isoform or promoter usage information because they sequence only the 3' end of the mRNA. By virtue of the higher cell count, these methods also tend to have lower read depths per cell, resulting in fewer genes (or transcripts) detected. Overall, the tradeoffs between full-length and 3'-end methods set up a choice: more in-depth information for few cells, or lower-resolution information for a greater number.

Recent studies have combined full-length scRNA library preparation with long-read sequencing to assay isoforms on the single-cell level. These have typically employed a hybrid approach, using short reads to identify cell types and quantify gene expression,

while using long reads to catalogue isoforms. The first of these studies were conducted on a very small number of cells (< 10)[75,76]. For instance, Byrne *et al.* applied Oxford Nanopore cDNA sequencing to seven mouse B1a cells and were able to detect many of the same genes captured in corresponding short-read data[76]. Overall, the sequencing depth was high but inconsistent, ranging from 17,749 to 128,726 Nanopore reads per cell. Gupta *et al.* took an different approach, developing the ScISOr-seq method to profile a higher number of cells (>1,000) from the mouse brain at the expense of low read depth (median 270 long reads per cell)[77]. In ScISOr-seq, short-read 3'-end sequencing is used to assign a cell type identity to each cell, and long-read PacBio sequencing is used to call isoforms. This hybrid approach has made it possible to examine alternative isoform usage across different cell types for hundreds of cells. Similar approaches have been pioneered using Oxford Nanopore sequencing as well[53,78]. Notably, in all of these studies, the read counts per cell were too low or too variable to attempt direct quantification of isoforms from the long reads themselves. To the best of our knowledge, no publication has achieved this yet.

As the field moves toward isoform-level analyses in single cells, it will have to contend with many of the same challenges faced by conventional scRNA-seq on the gene level. These include amplification bias, signal dropout, and the delicate matter of distinguishing between true biological variation and technical artifacts[79]. Compared to bulk methods, single-cell experiments must work with much smaller starting amounts of RNA, requiring multiple cycles of amplification that may distort the underlying proportions of the transcripts[80]. Runaway amplification events can be addressed by using unique molecular identifiers (UMIs) to identify and remove PCR duplicates[81]. However, UMIs

cannot address PCR biases themselves, nor the underlying sampling issues faced by single-cell sequencing. For instance, signal dropout is a persistent challenge for scRNA-seq. When the abundance of a particular gene is recorded as zero for a cell, it could be a true null value or simply a false negative resulting from low read depth and inefficient RNA capture[79,80]. Unfortunately, this is likely to be an even bigger problem for isoforms since individual transcripts are per definition expressed at a level less than or equal to the genes that they originate from. Additionally, the differences observed in single-cell transcriptomics can arise from many sources, including biological factors such as cell type and cell cycle stage, but also from technical sources such as library preparation and sequencing batch effects. Single-cell analysis suites such as Seurat, Scanpy, and MAST have all implemented methods to regress out noise from technical sources and to deal with signal dropout, but these are ongoing challenges[82–84]. Isoform quantification on the single-cell level necessitates the combination of deep long-read sequencing with improved computational methods for analysis.

1.3. Conclusions

Alternative splicing of mRNA allows the same gene to give rise to different isoforms. These may have distinct functional properties and are often associated with specific tissues and developmental contexts. Disruptions to the splicing process or misexpression of an isoform have been implicated in a variety of diseases, particularly neurological disorders. Therefore, understanding isoform-level gene expression in biological systems of interest is crucial. Short-read RNA-seq, the most commonly applied RNA sequencing technology, lacks isoform resolution due to the short length of the reads relative to the typical mammalian

mRNA transcript. Emerging long-read sequencing technologies overcome this limitation, but new computational methods are needed to address the unique challenges associated with these techniques. With this in mind, the overall goal of my work has been to develop methods for long read transcriptome analyses that are technology-agnostic with respect to platform and that directly use long reads for gene and transcript-level quantification.

In **Chapter 2**, I discuss TranscriptClean, a variant-aware, reference-based approach to correcting common sequencing errors in long reads. I applied TranscriptClean to publicly available PacBio data and showed that the method was able to salvage many reads that originally contained artifactual noncanonical splice junctions, allowing the corrected data to be used in downstream analyses. The results of this study were published in the journal *Bioinformatics* in January 2019.

Next, in **Chapter 3**, I describe TALON, a pipeline I developed for the ENCODE consortium to annotate, quantify, and filter long read transcripts. In collaboration with my co-first author Gabriela Balderrama-Gutierrez, Fairlie Reese, and others, we applied TALON to the transcriptome of human cell line GM12878 sequenced on both the PacBio and the direct-RNA Oxford Nanopore platforms. We found that PacBio more reliably captured full-length mRNA transcripts, but that it was prone to reverse-transcription artifacts. Additionally, we compared long-read PacBio transcriptomes from the mouse cortex and hippocampus, identifying both known and novel isoforms specifically expressed in each. This manuscript was initially posted on BioRxiv in pre-print form in June 2019 and was updated in March 2020. It is currently under review.

In **Chapter 4**, I present a collaboration with Barbara Wold's laboratory at Caltech to extend the TALON pipeline to single-cell PacBio data from the developing mouse limb bud. To focus our efforts on specific cell types of interest, we first used short-read RNA-seq (SMART-seq) as a screening tool to distinguish cells from three different muscle stages as well as the erythro-myeloid progenitors and tissue-resident macrophages. The selected cells were then sequenced using the PacBio platform. Focusing on a relatively smaller number of cells allowed us to achieve deep long-read coverage, enabling detection of quantitative gene and isoform differences between cell types and differentiation stages. My role in the project has been to develop computational analyses and methods, while the experimental work was performed by my collaborators, notably Dr. Brian Williams at Caltech.

Finally, in **Chapter 5**, I propose possible extensions of this work. Long-read transcriptomics is expanding rapidly as a field with a variety of applications to explore. For instance, long-read platforms are suitable for measuring allele-specific isoform expression provided that the data can be appropriately phased. In addition, I discuss some of the ongoing challenges posed by long-read artifacts. Overall, this thesis introduces and demonstrates the use of novel long-read transcriptome analysis methods on both bulk and single-cell RNA-seq samples.

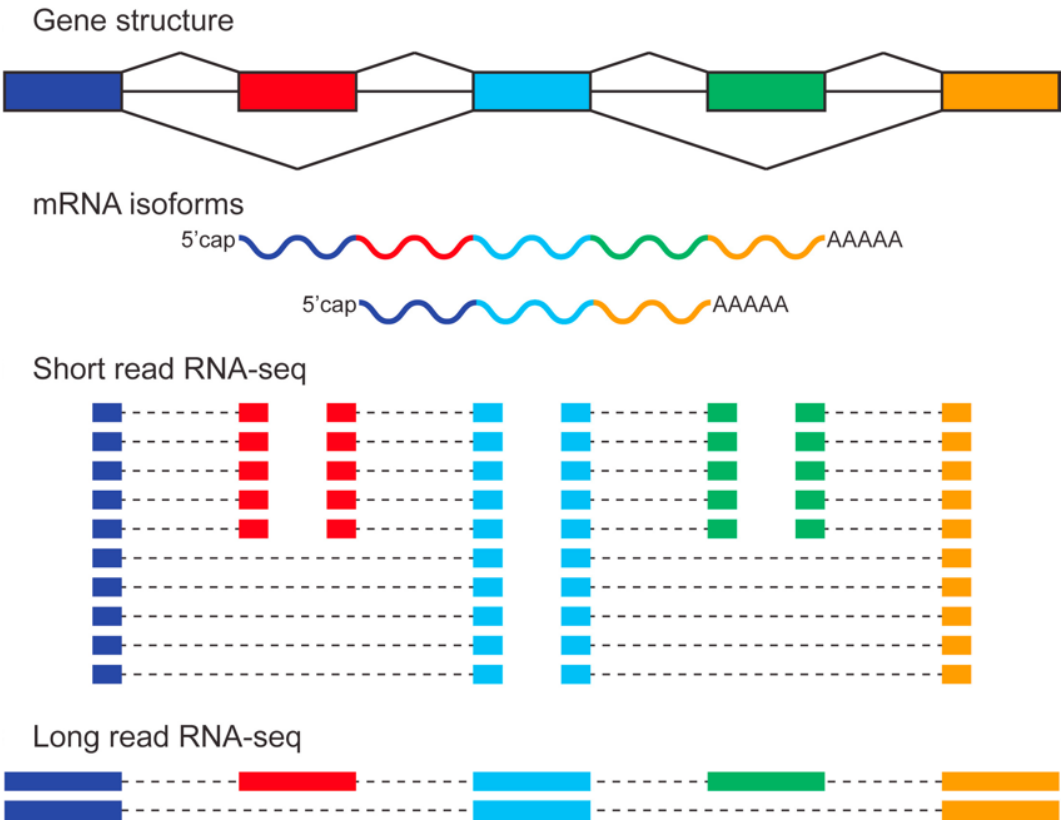

Figure 1.1. Isoform sequencing is straightforward using long reads. Short reads can capture individual splice junctions and exons but cannot recapitulate whole transcripts (Figure from Park *et al.* 2018).

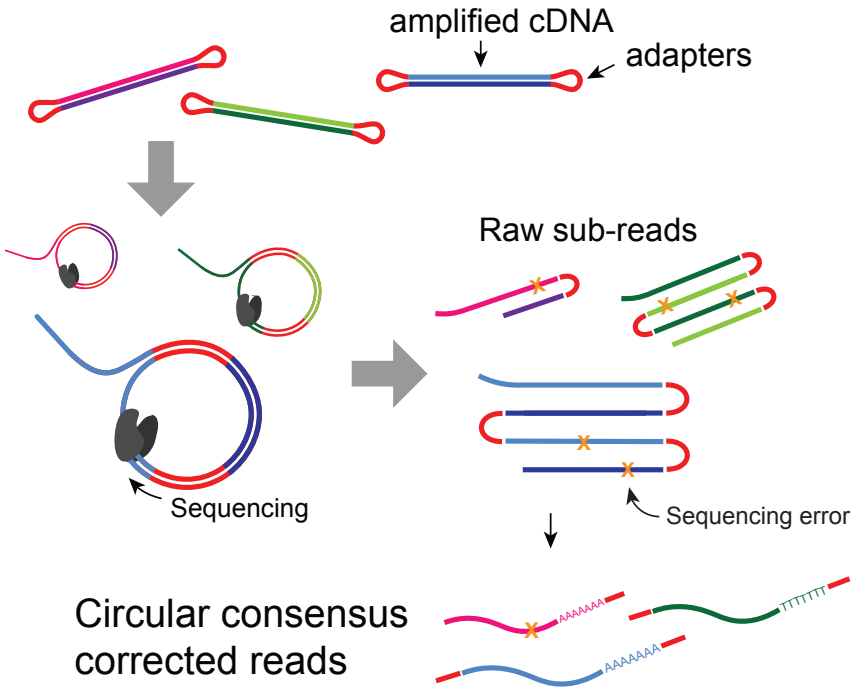

Figure 1.2. Circular consensus sequencing improves PacBio accuracy. Prior to PacBio sequencing, circular adaptors are attached to each end of the double-stranded cDNA. During sequencing, the template opens up into a circle, allowing the DNA polymerase to sequence multiple passes over the insert. These passes, or sub-reads, can be identified based on the spacing of the adaptor sequences. Consensus correction is performed by computationally comparing and merging the subreads for each molecule obtain the sequence that agrees best across the passes.

1.4. References

1. Uzman, A., Lodish, H., Berk, A., Zipursky, L. & Baltimore, D. Molecular Cell Biology (4th edition) New York, NY, 2000, ISBN 0-7167-3136-3. *Biochem. Mol. Biol. Educ.* (2000). doi:10.1016/S1470-8175(01)00023-6
2. Early, P. *et al.* Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell* (1980). doi:10.1016/0092-8674(80)90617-0
3. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
4. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.* (2003). doi:10.1146/annurev.biochem.72.121801.161720
5. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: Towards a cellular code. *Nature Reviews Molecular Cell Biology* (2005). doi:10.1038/nrm1645
6. Boise, L. H. *et al.* bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* (1993). doi:10.1016/0092-8674(93)90508-N
7. Takehara, T., Liu, X., Fujimoto, J., Friedman, S. L. & Takahashi, H. Expression and role of Bcl-xL in human hepatocellular carcinomas. *Hepatology* (2001). doi:10.1053/jhep.2001.25387
8. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**, 805–17 (2016).
9. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12**, 715–729 (2011).

10. Foulkes, N. S. & Sassone-Corsi, P. *More Is Better: Activators and Repressors from the Same Gene*. *Cell* **66**, (1992).

11. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* (2008). doi:10.1038/nature07509

12. Porter, R. S., Jaamour, F. & Iwase, S. Neuron-specific alternative splicing of transcriptional machineries: Implications for neurodevelopmental disorders. *Molecular and Cellular Neuroscience* (2018). doi:10.1016/j.mcn.2017.10.006

13. Weyn-Vanhentenryck, S. M. *et al.* Precise temporal regulation of alternative splicing during neural development. *Nat. Commun.* (2018). doi:10.1038/s41467-018-04559-0

14. Ule, J. *et al.* Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* (2005). doi:10.1038/ng1610

15. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol.* **5**, R74 (2004).

16. Love, J. E., Hayden, E. J. & Rohn, T. T. Alternative Splicing in Alzheimer's Disease. *J. Park. Dis. Alzheimer's Dis.* **2**, (2015).

17. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U. S. A.* (1977). doi:10.1073/pnas.74.12.5350

18. Alwine, J. C. *et al.* Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzyloxymethyl paper. *Methods Enzymol.* (1979). doi:10.1016/0076-6879(79)68017-5

19. Oberhauser, A. F., Balan, V., Fernandez-Badilla, C. L. & Fernandez, J. M. RT-PCR

cloning of Rab3 isoforms expressed in peritoneal mast cells. *FEBS Lett.* (1994). doi:10.1016/0014-5793(94)80409-5

20. Russell, H., Carita, F., Gavin, D. & Robert, W. Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Biotechnology* (1993).
21. Kafert, S., Krauter, J., Ganser, A. & Eder, M. Differential quantitation of alternatively spliced messenger RNAs using isoform-specific real-time RT-PCR. *Anal. Biochem.* (1999). doi:10.1006/abio.1999.4016
22. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature Genetics* (2002). doi:10.1038/ng0102-13
23. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* **102**, 11–26 (2018).
24. Johnson, J. M. *et al.* Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science (80-.).* (2003). doi:10.1126/science.1090100
25. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
26. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (80-.).* (2008). doi:10.1126/science.1158441
27. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
28. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Poznan, Poland)* **19**, A68-77

(2015).

29. Chang, J. T. H., Lee, Y. M. & Huang, R. S. The impact of the Cancer Genome Atlas on lung cancer. *Translational Research* (2015). doi:10.1016/j.trsl.2015.08.001

30. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* (2018). doi:10.1016/j.cell.2018.03.033

31. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).

32. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* (2012). doi:10.1101/gr.133744.111

33. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp120

34. Li, Y., Rao, X., Mattox, W. W., Amos, C. I. & Liu, B. RNA-seq analysis of differential splice junction usage and intron retentions by DEXSeq. *PLoS One* (2015). doi:10.1371/journal.pone.0136653

35. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

36. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, (2016).

37. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5593-601 (2014).

38. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

39. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* **323**, 133–138 (2009).

40. Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics. Proteomics Bioinformatics* **13**, 4–16 (2015).

41. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* (2015). doi:10.1016/j.gpb.2015.08.002

42. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* (2012). doi:10.1038/nbt.2280

43. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* (2016). doi:10.1038/nbt.3423

44. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* (2019). doi:10.1186/s13059-019-1727-y

45. Timp, W., Comer, J. & Aksimentiev, A. DNA base-calling from a nanopore using a viterbi algorithm. *Biophys. J.* (2012). doi:10.1016/j.bpj.2012.04.009

46. Schreiber, J. & Karplus, K. Analysis of nanopore data using hidden Markov models. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv046

47. Krishnakumar, R. *et al.* Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Sci. Rep.* (2018). doi:10.1038/s41598-018-21484-w

48. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).

49. Pomerantz, A. *et al.* Real-time DNA barcoding in a rainforest using nanopore

sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* (2018). doi:10.1093/gigascience/giy033

50. Castro-Wallace, S. L. *et al.* Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Sci. Rep.* (2017). doi:10.1038/s41598-017-18364-0

51. Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* (2010). doi:10.1093/nar/gkq543

52. Gordon, S. P. *et al.* Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PLoS One* **10**, e0132628 (2015).

53. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* (2018). doi:10.1073/pnas.1806447115

54. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci.* (2014). doi:10.1073/pnas.1400447111

55. Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708 (2016).

56. Abdel-Ghany, S. E. *et al.* A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* **7**, 11706 (2016).

57. Sahlin, K., Tomaszkiewicz, M., Makova, K. D. & Medvedev, P. Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nat. Commun.* **9**, 4601 (2018).

58. Good, B. H., De Montjoye, Y. A. & Clauset, A. Performance of modularity maximization

in practical contexts. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* (2010). doi:10.1103/PhysRevE.81.046106

59. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* (2020). doi:10.1186/s13059-020-1935-5

60. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* (2018). doi:10.1101/gr.222976.117

61. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* (2020). doi:10.1038/s41467-020-15171-6

62. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* (2019). doi:10.1186/s13059-019-1910-1

63. Altschuler, S. J. & Wu, L. F. Cellular Heterogeneity: Do Differences Make a Difference? *Cell* (2010). doi:10.1016/j.cell.2010.04.033

64. Goldman, S. L. *et al.* The impact of heterogeneity on single-cell sequencing. *Frontiers in Genetics* (2019). doi:10.3389/fgene.2019.00008

65. Durlanik, S. & Thiel, A. Requirement of immune system heterogeneity for protective immunity. *Vaccine* (2015). doi:10.1016/j.vaccine.2015.05.096

66. Guo, G. *et al.* Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Dev. Cell* (2010). doi:10.1016/j.devcel.2010.02.012

67. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

68. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* (2015). doi:10.1016/j.cell.2015.05.002

69. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* (2011). doi:10.1038/nbt.2038

70. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (80-.).* (2018). doi:10.1126/science.aam8999

71. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* (2017). doi:10.1038/ncomms14049

72. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* (2011). doi:10.1101/gr.110882.110

73. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: From vision to reality. *Nature* (2017). doi:10.1038/550451a

74. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* (2018). doi:10.1038/s41586-018-0590-4

75. Karlsson, K. & Linnarsson, S. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics* (2017). doi:10.1186/s12864-017-3528-6

76. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* (2017). doi:10.1038/ncomms16027

77. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).

78. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput, error corrected Nanopore single cell transcriptome sequencing. *bioRxiv* (2019).

doi:10.1101/831495

79. Yuan, G. C. *et al.* Challenges and emerging directions in single-cell analysis. *Genome Biology* (2017). doi:10.1186/s13059-017-1218-y

80. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* (2015). doi:10.1038/nrg3833

81. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* (2014). doi:10.1038/nmeth.2772

82. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3192

83. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* (2018). doi:10.1186/s13059-017-1382-0

84. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* (2015). doi:10.1186/s13059-015-0844-5

CHAPTER 2

TranscriptClean: A reference-based, variant-aware method for correcting sequencing errors in long reads

Note: This chapter was published in *Bioinformatics* in January 2019.

Chapter 2

TranscriptClean: A reference-based, variant-aware method for correcting sequencing errors in long reads

2.1 Abstract

Long-read, single-molecule sequencing platforms hold great potential for isoform discovery and characterization of multi-exon transcripts. However, their high error rates are an obstacle to distinguishing novel transcript isoforms from sequencing artifacts. Therefore, we developed the package TranscriptClean to correct mismatches, microindels and noncanonical splice junctions in mapped transcripts using the reference genome while preserving known variants. Our method corrects nearly all mismatches and indels present in a publicly available human PacBio Iso-seq dataset, and rescues 39% of noncanonical splice junctions.

2.2 Introduction

Conventional short-read RNA sequencing is widely used to quantify gene expression in a variety of applications. While cost-effective and accurate, short reads lack the ability to resolve full-length mammalian isoforms, which are commonly multiple kilobases long[1]. Long-read sequencing platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore bypass the transcript reconstruction challenges of short reads but have substantially higher error rates. Raw PacBio reads have a stochastic error rate of 11-15%, including single-base mismatches and microindel errors[2]. Microindels are especially problematic during isoform mapping because they can misrepresent splice junction locations.

Circular consensus correction and read polishing steps in the PacBio ToFU analysis pipeline can substantially reduce the error rate for most transcripts once raw reads are processed[2,3]. However, this correction process is only effective when multiple sequencing passes over the same insert molecule are available, which becomes less likely as transcript length increases[4].

To address this problem, various PacBio-specific tools have been developed to correct transcripts downstream of the ToFU pipeline. TAPIS, HapIso, and SQANTI use a reference-guided approach to correct indels within exons[5–7]. HapIso distinguishes single nucleotide variants from errors in a haplotype-aware manner by phasing long reads. TAPIS and SQANTI deal with remaining errors by removing affected transcripts, the former using a splice junction quality filter, and the latter using a random forest classifier. While these methods produce cleaner PacBio datasets, none of them attempt to correct noncanonical splice junctions arising from microindel errors. Furthermore, HapIso requires multiple transcripts per gene in order for the phasing to work, which is not a given depending on sequencing depth and gene expression level.

We present TranscriptClean, a program that uses the reference genome, splice annotation, and a variant file to correct mismatches, microindels, and noncanonical splice junctions in PacBio transcripts while preserving known variants. Running TranscriptClean on a publicly available PacBio human transcriptome from GM12878[8], we corrected 99% of indels, 98% of mismatches, and 39% of noncanonical splice junctions present in these

transcripts. This allowed us to salvage 32,536 transcripts that would have been discarded under previous workflows because of noncanonical splice junctions.

2.3 Methods

2.3.1 Indel and mismatch correction

TranscriptClean processes transcripts in the SAM format, scanning each entry to look for insertions, deletions, and mismatches relative to the reference genome. Indels less than or equal to the size threshold (default ≤ 5 bp) are modified to match the reference sequence. Mismatches in the transcripts are replaced with the reference base. Indel and mismatch correction can also be run in variant-aware mode to avoid removing variants of interest to the user. In this mode, mismatches and indels are changed to the reference sequence only if they do not match the position and sequence of a known variant in a user-provided VCF file. A potential downside of running mismatch correction is that it will remove novel SNPs or RNA editing events not provided in the VCF.

TranscriptClean outputs a SAM file of corrected transcripts with updated CIGAR, sequence, and MD/NM fields. It also provides a fasta file of corrected sequences alongside log files tracking changes to individual errors and transcripts. The accessory script generate_report.R produces figures summarizing the TranscriptClean results, and can also be used to choose an appropriate indel size threshold for a given dataset, as the size distribution may vary across different PacBio chemistries.

2.3.2 Noncanonical splice junction correction

TranscriptClean also provides the option of correcting noncanonical splice junctions. During pre-mRNA splicing, dinucleotides at the start and end of the intron form highly conserved canonical motifs GTAG, GCAG, and ATAC, with GTAG accounting for 98.9% of known human splice junctions[9,10]. Noncanonical splice junctions (NCSJs) are very rare events, which suggests that most NCSJs in long-read transcripts are likely to be sequencing errors. Typically 10-20% of PacBio transcripts contain at least one NCSJ[7].

When a microindel error disrupts a splice boundary, the read mapping can be affected in a variety of ways. In one scenario, the entire junction is shifted upstream or downstream of its original location. In another, the error is split across the junction, resulting in a smaller indel on each side. Finally, the error may only affect one side of the junction.

To identify NCSJs, TranscriptClean checks the intron motif of each transcript splice site. Each NCSJ is compared to user-provided high-confidence splice junctions (derived from same-sample mapped short RNA-seq reads or a reference annotation) and is changed to match the known junction when the distance between the NCSJ and its nearest high-confidence junction is microindel-sized.

2.4 Results and Discussion

We performed two TranscriptClean runs on ToFU-processed circular consensus GM12878 PacBio transcripts from Tilgner 2014 **(Table 2.1)**. In the first, we used known human splice junction annotations from GENCODE v24 and no variant file. Next we provided GM12878-

specific variants and splice junctions derived from GM12878 short reads. When provided GM12878-specific references for correction, TranscriptClean corrected 99% of indels and 39% of NCSJs, rescuing 32,536 transcripts no longer considered noncanonical. 98% of mismatches were corrected, with the remaining 2% representing known NA12878 SNPs.

A major goal of long-read isoform characterization is to provide a higher-quality reference transcriptome for short-read quantitation. If such a reference contains frequent sequencing errors, reads will not map well to it, defeating its purpose. Furthermore, downstream analysis programs commonly ignore transcripts with one or more NCSJs, effectively throwing out long-read data that could provide interesting isoform information. Repairing errors where possible allows more data to be used, particularly for longer transcripts. While the current version of variant-aware TranscriptClean does not account for the case of a sequencing error converting a real SNP to the reference base, nor the case where a real indel is disguised by one or more sequencing errors, we hope to improve correction for special cases like these in future versions and to support transcript correction of Oxford Nanopore reads.

Table 2.1. Summary of GM12878 TranscriptClean results

	No TC	TC with GENCODE splice junctions	Corrected	TC with GM12878 Illumina SJs & variants	Corrected
Total Transcripts	568048	568048	---	568048	---
Canonical Transcripts	479005	512092	---	511541	---
Noncanon.Transcripts	89043	55956	37%	56507	37%
Deletions	3133172	23047	99%	29883	99%
Insertions	1901787	20175	99%	21816	99%
Mismatches	14380068	0	100%	295547	98%
NCSJ	109268	66304	39%	66784	39%

2.5 References

1. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
2. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* **323**, 133–138 (2009).
3. Gordon, S. P. *et al.* Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PLoS One* **10**, e0132628 (2015).
4. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* (2015).
5. Abdel-Ghany, S. E. *et al.* A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* **7**, 11706 (2016).
6. Mangul, S. *et al.* HapIso: An Accurate Method for the Haplotype- Specific Isoforms Reconstruction From Long Single-Molecule Reads. *IEEE Trans. Nanobioscience* **16**, 108–115 (2017).
7. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* (2018).
8. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci.* (2014).
9. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
10. Parada, G. E., Munita, R., Cerda, C. A. & Gysling, K. A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.* (2014).

CHAPTER 3

A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification

Note: Gabriela Balderrama-Gutierrez and I contributed equally to the material in this chapter. I developed the computational methods and designed the human cell line analyses, whereas she generated the long-read datasets and designed the mouse brain analyses. The chapter has been posted as a preprint on BioRxiv (https://doi.org/10.1101/672931), and is currently under review.

Chapter 3

A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification

3.1 Abstract

Alternative splicing is widely acknowledged to be a crucial regulator of gene expression and is a key contributor to both normal developmental processes and disease states. While cost-effective and accurate for quantification, short-read RNA-seq lacks the ability to resolve full-length transcript isoforms despite increasingly sophisticated computational methods. Long-read sequencing platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) bypass the transcript reconstruction challenges of short reads. Here we introduce TALON, the ENCODE4 pipeline for platform-independent analysis of long-read transcriptomes. We apply TALON to the GM12878 cell line and show that while both PacBio and ONT technologies perform well at full-transcript discovery and quantification, each displayed distinct technical artifacts. We further apply TALON to mouse hippocampus and cortex transcriptomes and find that 422 genes found in these regions have more reads associated with novel isoforms than with annotated ones. We demonstrate that TALON is a capable of tracking both known and novel transcript models as well as their expression levels across datasets for both simple studies and in larger projects. These properties will enable TALON users to move beyond the limitations of short-read data to perform isoform discovery and quantification in a uniform manner on existing and future long-read platforms.

3.2 Introduction

Differences in gene expression play a large role in shaping cell phenotypes and interactions, both during development and in later life. While humans have around 20,000 protein coding genes, they produce at least 100,000 splice isoforms through alternative splicing, and potentially many more[1]. Alternative splicing controls which exons are included in the mature mRNA, thus expanding the number of possible transcripts that a single gene can encode. Some isoforms have vastly different functions and may be highly specific to a particular tissue or temporal stage[2–4]. For instance, alternative splicing of the transcription factor *erbAα* in rats gives rise to one isoform which acts as a transcriptional activator, while a second isoform acts as a repressor[5]. This is a specific instance of an evolutionary strategy whose extent is not yet known, in which differential RNA splicing creates one or more "dominant negative" protein isoforms. Differential RNA isoforms are also important in disease. The *Mapt* gene has isoforms that are known to be differentially expressed in various human neural lineages, and their relative proportions change during progression of Alzheimer's disease, ultimately leading to the formation of the tangles that kill neurons[6].

In the best understood cases, alternative splicing is tightly regulated, relying on highly conserved sequence and structure motifs and complex networks of RNA binding protein interactions to define functional isoforms[7]. Disruptions to the splicing process frequently lead to disease, whether in the form of genetic mutations that directly affect splice sites or splicing factors, or more subtle changes that alter the balance between different isoforms[6,7]. As a result, alternative splicing and exon usage in RNA transcripts have long been the subject of great interest in the context of development and disease. In early studies, the preferred

methods for characterizing and measuring isoforms were RT-PCR, Sanger sequencing of expressed sequence tags (ESTs), and isoform-specific microarrays[8]. This changed dramatically with the availability of next-generation short-read RNA sequencing, which allows gene expression to be profiled quantitatively in a high-throughput manner[9]. This led to the generation of large reference transcriptome databases for human and mouse cell types and tissues, beginning with ENCODE and rapidly expanding to GTEx and FANTOM[10–12]. In the cancer community, the Cancer Genome Atlas (TCGA) serves as a massive source of RNA-seq data from patient samples[13].

With the widespread availability of RNA-seq, many efforts have been made to infer isoform usage from short-read data[14]. However, this is intrinsically challenging, as short-read protocols require cDNA transcripts to be sheared into 50-300 bp pieces prior to sequencing. These pieces are far smaller than typical mammalian transcripts, which can be multiple kilobases in length[15]. This means that it is not possible to know the exact combination of exons originally present in each transcript molecule. To get around this, computational methods were developed to reconstruct the transcript models present in a sample and to quantify their abundance. Here, we use the term 'transcript model' to describe a distinct set of splice junctions paired with variable 5' and 3' ends. Bioinformatics software packages such as Kallisto use expectation-maximization to pseudo-align short reads to a transcriptome reference, generating abundance estimates for transcript and gene models[16]. These algorithms are effective in broadly identifying which transcripts the reads are compatible with, but they cannot tell exactly which ones were present. Long-distance contiguity is especially challenging. An additional drawback is that these methods depend

heavily on the choice of the reference transcript annotation and, as such, they cannot identify novel transcript models. Another widely used approach to quantifying alternative splicing is to compute short read coverage of specific splice junctions or exons, and compare the resulting counts across samples using statistical tests[17,18]. While these methods are useful for detecting alternative exon usage, they do not overcome the fundamental limitations of short-read data with respect to assembling and assigning exactly which exons made up the source transcript.

Since 2010, third-generation sequencing platforms such as PacBio and Oxford Nanopore (ONT) have pioneered the use of long reads in genomics[19,20]. With read lengths of up to 60 kb for PacBio and up to 1 Mb for Oxford Nanopore, these reads can capture entire transcripts from end to end. They also offer the advantage of representing single molecules rather than amplified clusters, making them ideal for sequencing isoforms. Historically, the major drawbacks of long read technologies have been their relatively low throughput as well as high indel and mismatch error rates ranging up to 15-20%[19]. In the case of PacBio, these stochastic errors are mitigated by using circular consensus sequencing, in which multiple sequencing passes over the same molecule are used for error correction[21]. The exact error rate depends largely on the number of passes that a molecule receives. Computational methods have also been developed to correct errors in long reads, including hybrid approaches that incorporate short reads, and other methods that make use of reference annotations[22–25].

Due to the low throughput of the original platforms, the conventional long-read transcriptome sequencing approach was to first catalog expressed isoforms using long reads from size-selected subsamples, and then map short reads to the resulting transcriptome references for the purpose of quantification[26–28]. PacBio popularized this method in mammals, plants, and beyond under the name "Iso-seq". Recently, PacBio yields increased substantially, producing up to 8 million reads per SMRT cell on the Sequel 2 compared to 150,000 on the older RSII machines. Similar yield increases have been reported for Oxford Nanopore. This increased throughput has made direct long-read quantification more plausible. Unfortunately, most existing tools for analyzing long-read transcriptome data were not explicitly designed for this purpose. PacBio-affiliated software packages such as ICE-Quiver/Arrow and Cupcake ToFU generate *de novo* transcript models by clustering long reads and then merging them to generate one transcript model per cluster[26,29]. This is a particularly useful approach in species that lack a reference genome, but it comes with disadvantages. ICE-Quiver has been known to merge together transcripts from highly similar genes and can smooth over real differences of interest such as sequence variants and RNA editing events[30]. In addition, the algorithm is stochastic by nature, and cluster assignments for individual reads can vary substantially across different runs. Most existing programs for transcriptome-wide PacBio annotation and quantification rely on the ICE-Quiver or Cupcake ToFU outputs. For instance, SQANTI uses post-ToFU transcript models and their estimated abundances as the input to its annotation, quantification, and quality control pipeline[23]. Another set of pipelines such as FLAIR have been developed for analyzing Oxford Nanopore cDNA and direct RNA sequencing data[31]. As in ICE-Quiver, a common feature of these

pipelines is the alignment of reads to each other before determining which known and novel transcripts are present.

Here, we present TALON, the ENCODE4 pipeline for simultaneous transcript discovery and quantification of long-read RNA-seq data regardless of platform. This pipeline is designed to explicitly track both known and novel transcripts across different bio-samples to allow for annotation and use of new isoforms. The full TALON pipeline is available on GitHub through the ENCODE4 Data Coordinating Center (DCC) at ENCODE-DCC/long-read-rna-pipeline and at mortazavilab/TALON. We first analyze the transcriptome of the GM12878 cell line using the PacBio and ONT to quantify the relative performance of both platforms. The TALON pipeline allows us to process PacBio and ONT data in a uniform fashion and make direct comparisons between the two. We evaluate the resulting transcriptomes relative to available CAGE, poly(A), and RNA-PET annotations in these cells and find that each long-read technology is affected by different artifacts. We then sequence the transcriptomes of adult mouse hippocampus and cortex to show the applicability of the TALON pipeline for the analysis of complex tissues. Overall, we demonstrate that current long-read platforms are suitable for quantifying and characterizing isoform-level expression of genes.

3.3 Results

3.3.1 Tracking transcript novelty and quantification using TALON

To compare long read platforms side by side and to track isoforms consistently across multiple datasets, we developed a technology-agnostic long-read pipeline called TALON **(Figure 3.1a)**. This pipeline is designed to annotate full-length reads as known or novel transcripts and also to report the abundance for genes and transcripts across datasets. Starting from long reads mapped to the reference genome with a long-read aligner such as Minimap2, reference-based error correction is performed using TranscriptClean to remove microindels, mismatches, and noncanonical splice junctions in a variant-aware manner as previously described[25]. Noncanonical splice junctions are permitted in the final output only if they are supported by the splice annotation. Note that TALON expects reads to be oriented to the appropriate strand, which is typically achieved using platform-specific preprocessing in the case of cDNA reads **(Figure S3.1a-b).** After TranscriptClean, corrected reads are passed into the *talon_label_reads* TALON module, which records QC information for use by subsequent steps. In particular, long-read libraries built using poly(A) selection are prone to internal priming artifacts in A-rich regions of transcripts that result in truncated isoforms. Therefore, tracking the fraction of As following alignments is informative for TALON's transcript filtering process. After the internal priming labels have been assigned, the reads are passed into the main *talon* module for annotation. In a talon run, each input SAM read is compared to known and previously observed novel transcript models on the basis of its splice junctions, start, and end points. This allows us to not only assign a novel gene or transcript identity where appropriate, but to track new transcript models and characterize how they differ from known ones. The result is a collection of all transcripts observed in each input dataset that can then be filtered, quantified, and compared using downstream TALON modules.

We adopted the nomenclature introduced by SQANTI to characterize the different types of transcript novelty in our datasets[23]. Query transcripts with splice junctions that perfectly match an existing model are deemed 'known' **(Figure 3.1b)**. Flexibility is allowed at the 5' and 3' ends. In cases where a transcript matches a subsection of a known transcript model and has a novel putative start or endpoint, it is considered an 'incomplete splice match' (ISM). TALON further subdivides the ISM category into prefix ISMs and suffix ISMs. The former refers to ISMs that match along the 5' end of an existing transcript model, and the latter describes ISMs that match to the 3' end. It is possible for a transcript to belong to more than one ISM category if it matches to different parts of several existing transcript models. The ISM category is useful as a means of quality control as libraries with a higher proportion of ISMs relative to known transcripts tend to be less than complete in terms of length and may harbor more artifacts. For instance, RNA degradation and incomplete reverse-transcription can lead to suffix ISMs. In Oxford Nanopore, pore blockages can produce suffix ISMs by prematurely stopping sequencing of the RNA. In the case of prefix ISMs, internal priming is the most likely culprit. However, not all ISMs are sequencing artifacts. To differentiate between a truly novel ISM transcript and one that is artifactual, it is useful to test against relevant orthogonal data such as CAGE, RNA-PET, or poly(A) annotations, which are often available from external databases. This can provide independent validation to support or reject a new 5' or 3' end seen in an ISM transcript.

The next category, novel in catalog (NIC), describes transcripts that have known splice donors and acceptors, but reveal new connections between them. This can be thought of as a novel arrangement of known exons. Novel not in catalog (NNC) transcripts contain at

least one novel splice donor or acceptor, meaning that there is at least one novel exon boundary present. Genomic transcripts are either partial transcripts that do not share any splice junctions with overlapping genes or may come from DNA contamination in the samples, and are therefore discarded by the filter, reproducible or not. The antisense category consists of transcripts that overlap an existing gene, but are oriented in the opposite direction. If a transcript lacks any overlap with a known gene, then it is deemed intergenic. Taken together, the novelty categories allow us to examine the types of transcripts that we detect in our long-read datasets, to perform quality control, and to stratify or filter by category.

Biological replicates serve as an important means of verifying novel transcript discoveries. Although the accuracies of long-read platforms are improving, artifactual transcripts are still a problem, and may arise from a variety of technical sources. TALON streamlines the filtering process for multiple datasets by tracking transcript annotations and abundance in one place, where the information can be easily accessed and compared. Our filtering process uses the novelty labels assigned to each observed transcript model in order to remove likely artifacts. Observed transcripts that fully match counterparts in the GENCODE annotation are accepted immediately, but we require that novel transcripts be supported by at least 5 reads each in at least two biological replicate samples in order to be included in the downstream analysis. Furthermore, all five reads must all pass the internal priming cutoff (fraction As ≤ 0.5). These cutoffs can be adjusted by the user to accommodate different oligo-dT lengths or sequencing depths. As additional samples are sequenced, it is also possible to cross-reference novel transcripts across these datasets.

TALON quantification relies on the premise that each long read represents an individual transcript molecule sequenced. This allows us to quantify expression by simply counting the number of individual reads that were assigned to a particular transcript or gene and then converting these values into units of transcripts per million (TPM) to adjust for library size. For gene-level expression values, we include all reads assigned to a locus in the computation, since even incomplete transcripts (ISMs) that did not meet the threshold to become a new transcript model are informative for the overall gene expression level. On the transcript level, however, we apply the TALON filters in order to avoid quantifying transcript models with insufficient evidence.

To demonstrate the utility of TALON, we applied it in two different settings **(Table S3.1)**. First, we compared long-read GM12878 data sequenced on different platforms: PacBio Sequel II and direct-RNA ONT **(Figure 3.1c).** Then, we used TALON to analyze gene and isoform-level expression across the complex tissues of cortex and hippocampus in mouse **(Figure 3.1d)**. In each case, we sequenced at least 6 million raw reads per replicate. Spike-in RNA variants (SIRVs) in our samples provided us with an opportunity to evaluate TALON filtering on artificial sequences with fully known splice patterns. The expected outcome in an error-free setting would be to detect exactly 69 known isoforms from a total of 7 SIRV genes, and to detect zero novelty. After applying the TALON transcript filter (including the internal priming cutoff) to SIRVs sequenced with the two PacBio GM12878 replicates, we detected 67 known SIRV transcripts and only 13 novel models **(Figure S3.2a)**. 96% of the filtered reads matched a known isoform **(Figure S3.2b)**, In contrast, the unfiltered SIRV data contained a much higher fraction of artifactual novel transcripts **(Figure S3.2c,d)**. The

ISM category was the most common form of novelty, accounting for between 5 and 6% of the unfiltered reads by replicate. About 60% of the reads assigned to prefix ISMs displayed evidence of internal priming, suggesting that this is a substantial artifact of cDNA sequencing in PacBio **(Figure S3.2e)**. The TALON filter was highly effective in removing these transcripts- after filtering, only 9 ISM models remained. Overall, these results indicate that the TALON filter is effective at removing artifactual transcript models.

3.3.2 Performance of TALON on human ENCODE Tier 1 PacBio data

We then turned our attention to applying TALON to GM12878 reads mapped onto the human genome. TALON detected 15,727 known GENCODE genes and 26,841 GENCODE transcripts in GM12878 across the two replicates. The number of known genes is smaller than the number of known transcripts because known genes can be detected through novel transcripts as well as known ones. The analysis also called 359 unknown gene models, the majority of which consisted of monoexonic transcripts mapped as antisense within a known gene locus. The TALON N50 read lengths for Rep 1 and Rep 2 were 1,877 and 1,791 nucleotides, respectively, which is in line with the expected length distribution of most mammalian mRNA transcripts **(Fig S3)**.

We next computed the expression level for each known GENCODE gene across the PacBio data. For this quantification, we included all long reads assigned to a locus in these counts because even incomplete transcripts are informative for the overall gene expression level. The resulting gene expression levels were highly correlated across biological PacBio replicates of each cell line (Pearson r = 0.97, Spearman rho = 0.92) **(Figure 3.2a)**. This shows

that our PacBio primary data coupled with the TALON pipeline produces reproducible quantifications of gene expression.

We also compared our PacBio results to short-read RNA-seq data from the same cell line. First, we examined how often PacBio was able to detect known genes as a function of their short-read expression level **(Figure 3.2b)**. As expected, genes at the lower range of expression (< 2 TPM from short reads) were less likely to be detected by PacBio, but upwards of 70% of genes expressed above 2 TPM were reproducibly detected. Overall, the expression levels of the 14,947 genes detectable in both PacBio and Illumina correlated well across platforms (Spearman rho 0.78). We conducted a differential expression analysis to further examine how much gene expression levels vary between the platforms. The log fold change between PacBio and Illumina was computed using the exact test method in EdgeR, and Bonferroni correction for multiple testing was performed on the resulting p-values. This analysis revealed that there was no significant difference in expression levels for most genes **(Figure 3.2c).** However, a subset of genes showed significant fold change differences, including 773 that were higher in PacBio and 1,139 that were higher in Illumina. Genes expressed significantly higher in Illumina tended to have longer median transcript lengths on average than those that were not differentially expressed or that were expressed more highly in PacBio **(Figure S3.4a)**. This suggests that these PacBio data under-detect the longest transcripts (greater than 5 kb) when no size selection is applied. Genes with higher expression in PacBio had significantly higher median GC content as a group (adjusted p = 4.950e-08) than those that were higher in Illumina (**Figure S3.4b)**. It is possible that this is

related to the GC bias known to affect Illumina next-generation sequencing. Overall, non-size selected PacBio libraries detect most of the genes expressed at 1 or more TPM in Illumina.

Having established that TALON can quantify gene-level expression on the basis of long reads, we moved on to transcript-level quantification. As expected, most of the transcript models identified in our analysis of the extensively-studied GM12878 cell line were known matches to the GENCODE annotation **(Figure 3.2d)**. The expression levels of detected known transcripts were highly correlated across PacBio biological replicates (Pearson r = 0.97, Spearman rho = 0.73) **(Figure 3.2e)**. Novel transcript models displayed even stronger expression correlations, likely related to the stringent abundance and filtering requirements that were applied to them (Pearson r = 0.97, Spearman rho = 0.83) **(Figure 3.2f)**. PacBio transcript expression levels were not significantly different for 87% of GENCODE transcripts when compared to short-read expression levels **(Figure 3.2g)**. The known, NIC, and NNC isoform categories account for about 94% of the filtered PacBio reads, with known transcripts making up 91.1% of the reads **(Figure 3.2h)**. NIC and NNC transcripts contained a larger number of exons on average than the other novelty categories, and also tended to come from longer reads **(Figure S3.5a-b)**. To evaluate the canonical junctions found in the PacBio reads, we compared them to junctions called from the short-read Illumina GM12878 RNA-seq data using STAR[32]. 83% of novel PacBio splice junctions featuring known splice donors/acceptors had short-read support **(Figure 3.2i)**. The majority of PacBio junctions with a novel splice donor and/or acceptor were supported as well. Overall, these results indicate that we can reliably annotate and quantify transcript models using our long-read pipeline.

GM12878 is an Epstein-Barr Virus (EBV) transformed lymphoblastoid cell line (LCL). We were therefore able to analyze the gene and transcript expression of EBV within the GM12878 PacBio transcriptome. We found that EBV transcripts are detectable using long-read sequencing, and that these transcripts can be quantified, annotated, and assessed for their novelty using TALON. Overall, 25 known and 4 post-filter novel EBV transcript isoforms were detected and 28 known EBV genes were detected **(Figure S3.6a-b)**. Many detected transcripts belong to the *EBNA* gene family **(Figure S3.6c)**, which code for proteins that are essential to the virus' ability to transform infected cells into LCLs [33], and are typically among the most highly expressed genes from the EBV chromosome in LCLs.[34] Consistent with the novel transcript models detected by TALON, the *EBNA* transcripts have previously been identified as heavily alternatively spliced[35].

3.3.3 Performance of TALON on Oxford Nanopore data and comparison with PacBio

Oxford Nanopore is an alternative long-read sequencing platform that offers the option of direct RNA sequencing [36]. While the protocol involves one reverse-transcription step, this is primarily for the purpose of removing secondary RNA structure and ultimately only the RNA strand is sequenced. In order to demonstrate the applicability of TALON to the Nanopore platform, we directly sequenced RNA from two biological replicates of GM12878 to a depth of at least 2 million basecalled reads per replicate. After alignment with Minimap2[37], each replicate was processed with the TALON pipeline as described for PacBio. The TALON N50 read lengths for the datasets were 1,269 nucleotides for Rep 1 and 989 for Rep 2 **(Fig S7a-b)**. Although the starting number of reads was lower than in our PacBio

transcriptomes, we detected ~13,500 known GENCODE genes and ~18,000 known isoforms in GM12878. Gene and transcript expression levels across the two GM12878 ONT replicates correlated with each other (gene Pearson r = 0.99, gene Spearman rho = 0.92; known transcript Pearson r = 0.97, known transcript Spearman rho = 0.64) **(Figure 3.3a-b)**. When we labeled the transcripts by their novelty type, it became apparent that differences in isoform-level expression between ONT replicates are largely driven by overrepresentation of novel ISM transcript models **(Figure 3.3c-d)**. This leads us to believe that ONT is more sensitive to degradation events or is prone to stopping mid-transcript during sequencing, which may explain the high ISM numbers in our data.

Next, we compared gene and transcript expression levels across the PacBio and ONT platforms in GM12878 **(Figure 3.3e)**. These were well-correlated at the gene level, but there were interesting differences at the transcript level. For instance, ISMs were overrepresented in ONT relative to PacBio, suggesting that the former had more difficulty sequencing full-length transcripts (**Fig S3.7c**). On the other hand, of 414 total antisense transcripts called across the platforms, 276 were unique to PacBio, whereas only 26 were detected in ONT alone **(Figure 3.3f)**. This likely means that the majority of antisense transcripts were in fact artifacts of the reverse transcription steps required for PacBio, demonstrating a drawback of conversion to cDNA before sequencing, at least by the standard methods used for PacBio. Interestingly, there is a set of 88 genes with TPM > 10 in both technologies that are detected as more than 10-fold more highly expressed in Oxford Nanopore, which could represent further under-representation of these transcripts due to reverse transcription biases. Among the genes enriched in Oxford Nanopore we found a subset related to mitochondrial

functions (MT-RNR1, MTCO1, MT-CO2, MT-ATP6, MT-CO3 and MT-CYB), that have been previously characterized as a benchmark for direct RNA-seq performance as pointed out by other groups[31]. Although some mitochondrial genes are subject to a deadenylation process, mature mt-mRNA transcripts contain a non-templated sequence of poly(A)s[38]. This fact, along with the minimal processing steps before sequencing, might explain the higher detection levels of these genes on the Oxford Nanopore platform compared to PacBio.

3.3.4 Comparison of TALON and FLAIR on GM12878 PacBio and ONT data

FLAIR is another recent pipeline designed to identify and quantify transcripts in long-read PacBio or ONT data[39]. To compare FLAIR and TALON, we ran FLAIR on the full-length, non-chimeric PacBio reads from GM12878 replicates 1 and 2 as described in the Supplementary Methods. We then compared the FLAIR quantification results to those generated by TALON in Figure 3.2. Similarly to TALON, FLAIR reported strong gene and transcript-level expression correlations across biological replicates (FLAIR Pearson r = 0.96, Spearman rho = 0.94 for known genes and Pearson r = 0.96 Spearman rho = 0.88 for known transcripts in GM12878). However, FLAIR was less sensitive than TALON with respect to detecting known genes and transcripts **(Table S3.2)**. For instance, in GM12878, TALON detected 2,525 more GENCODE genes than FLAIR that were also expressed in the corresponding short-read data **(Figure S3.8a)**. Recognizing that FLAIR was initially developed for ONT data, we ran the same comparison on our direct-RNA ONT GM12878 datasets **(Fig. S3.8b)**. As in the PacBio analysis, FLAIR detected fewer known genes and transcripts in the ONT data than TALON **(Table S3.3)**. This discrepancy was particularly pronounced at lower expression levels, but applied to genes with > 50 TPM in Illumina as

well (**Fig. S8b).** Taken together, these results demonstrate that TALON is currently more sensitive to known genes and transcripts than FLAIR in the same datasets.

3.3.5 Assessing completeness of TALON transcript models using CAGE, poly(A) motifs, and RNA-PET

The exonuclease treatment of our samples at the RNA stage and the full-length classification step *in silico* are intended to ensure that the transcripts at the end of our pipeline have intact 5' and 3' ends. To verify completeness, we performed an integrative analysis comparing our TALON transcript models with data from the CAGE and RNA-PET assays, as well as computationally identified poly(A) motifs. For known transcript models, the annotated GENCODE 5' and 3' sites were used.

CAGE is a genome-wide method of annotating transcription start sites that works by trapping the 5' end cap of a mature mRNA transcript using an antibody and then sequencing its 5' end. To validate the 5' ends of our long-read transcript models, we compared them to CAGE-derived TSSs from the FANTOM5 project. 76% of known GENCODE transcripts in our GM12878 PacBio transcriptome had CAGE support **(Figure 3.4a)**. Transcripts in the prefix ISM category were overwhelmingly supported (97%), whereas suffix ISMs were not (34%). 94% of NIC and 87% of NNC transcripts were supported by CAGE, indicating that their 5' ends were at least as reliable as those of the known transcripts. However, the antisense PacBio transcripts had scant support, lending credence to the idea that they are largely reverse-transcription artifacts. We observed similar CAGE trends in our ONT transcriptome **(Figure 3.4b)**, although notably, most transcript categories tended to have lower rates of support than in the corresponding PacBio transcriptome.

To examine transcript completeness at the 3' end, we conducted a computational poly(A) motif analysis of our long-read transcript models. This entailed scanning the last 35 bases of each transcript sequence to look for the presence of a known poly(A) motif. In PacBio, 64% of known transcripts contained such a motif **(Figure 3.4c)**. Rates of support were also high in the suffix ISM, other ISM, NIC, and NNC categories (86%, 80%, 84%, and 86% respectively). As expected, only 43% of the prefix ISMs contained a poly(A) motif, indicating that many of these transcripts may be artifactual. Overall, similar trends were observed in the ONT transcripts **(Figure 3.4d)**.

Finally, we sought to validate the 5'-3' pairings in our transcript models using publicly available RNA-PET data from the ENCODE consortium for both PacBio and ONT transcriptomes **(Figure 3.4e-f)**. This assay marks the start and endpoints of individual cDNA transcripts by circularizing and sequencing them with paired-end tags. This data type was lower-throughput than the more recently generated CAGE data, which helps explain the lower rates of RNA-PET support for known transcripts. We nevertheless observed strong RNA-PET support for NIC and NNC transcripts in both PacBio and Oxford Nanopore. Of the three ISM categories, prefix ISMs were the most likely to have RNA-PET support for their 5'-3' end pairing. Antisense transcripts had extremely high rates of RNA-PET support. The RNA-PET protocol uses reverse transcription, and therefore it is possible that this assay is prone to the same types of antisense artifacts as PacBio.

Taken together, the results of our CAGE, poly(A), and RNA-PET analyses indicated that most NIC and NNC transcript models derived from long reads have intact 5' and 3' ends, which argues that they represent full-length RNA's. However, inferred transcripts in the ISM novelty category require more scrutiny. As expected based on the category definition, prefix ISMs had reliable 5' sites, but their 3' ends were potentially incomplete in many cases. The reverse was true for suffix ISMs. In both cases, this suggests that many are technical artifacts. In general, the PacBio platform did a better job of capturing complete transcripts in our hands than did direct-RNA ONT and offered the additional benefit of higher throughput.

3.3.6 Comparison of PacBio transcriptomes of mouse cortex and hippocampus

After testing and characterizing TALON on PacBio data in a homogeneous cell line, we applied it to begin to discover and quantify isoforms in the complex brain regions of the mouse cortex and hippocampus. The cortex and hippocampus are critical regions of the brain for learning because of their functions of neural integration and memory, respectively[41]. Therefore, these regions have been characterized extensively under different conditions and models in order to understand their gene expression profiles[42]. These brain regions are much more complex in cell type composition than isolated cell lines, and the two regions have both similar and distinct cell types. Regulation of cell type diversity is key during their development, aging, and in disease, with both known and likely undiscovered differences in gene and isoform-level expression[42]. In addition, the brain at large is known to have a high alternative splicing ratio when compared to other tissues [40].

We sequenced two PacBio Sequel II replicates each of cortex and hippocampus to a minimum depth of 6 million raw reads per replicate and ran TALON on them. Gene expression was highly correlated across biological replicates (cortex Pearson r = 0.96 , Spearman rho = 0.95 and hippocampus Pearson r = 0.89 Spearman rho = 0.83) as was transcript-level expression (cortex Spearman rho = 0.84 and hippocampus Spearman rho = 0.73) **(Fig. S3.9a-d)**. On average, we detected 17,000 known genes and 26,000 known transcripts for each tissue. The diversity of the isoform novelty categories was similar between cortex and hippocampus **(Figure 3.5a-b)**. We identified 694 differentially expressed transcripts isoforms from a total of 612 genes (log fold-change > 1 and adjusted pvalue < 0.01), including 607 known and 87 novel transcript models **(Figure 3.5c)**. This included differences between known transcripts for genes such as Pnisr, which is a splicing factor involved in aging[43]. Other examples involving novel isoforms include the maternally expressed 3 (Meg3) long-coding RNA gene for which we detected two NIC transcript models that were enriched in cortex. Meg3 is thought to be involved in controlling vascularization in the brain by inhibiting angiogenesis[44] and is highly expressed. In addition, an NNC transcript of Amigo2 was selectively enriched in hippocampus. Amigo2 is known to be upregulated in the CA2 and CA3a regions[45] and is also commonly used as a marker of astrocyte activation[46].

We extended our transcript analysis by asking whether there are specific sets of genes with more novel transcript expression than known transcript expression. Specifically, we focused on genes that had more reads assigned to NIC and NNC novel transcript isoforms than known transcript isoforms in either brain region and found a shared set of 352 genes with an additional 29 and 41 that were specific to cortex and hippocampus respectively

(Figure 3.5d). This includes an NIC isoform of Pnisr (ENCODEMT000904037) that is enriched in the cortex **(Figure 3.5e)**. These analyses show that full-length transcriptome sequencing can detect isoform differences even in intensively studied tissues and cell types. These differences would be difficult to recover from short reads alone, especially for the NIC isoforms.

3.4 Discussion

We demonstrated here that with sufficient sequencing depth, long read data can be used to reproducibly quantify gene and transcript expression in homogeneous cell lines and in complex tissues. Our technology-agnostic long-read pipeline, TALON, simplifies the process of comparing long-read transcriptomes across different datasets and allowed PacBio and ONT transcriptomes to be directly compared. We found that our PacBio results are reasonably well-correlated with Illumina short-read data, particularly for gene expression levels above 2 TPM. We also found that, in our hands, the current PacBio platform captured more complete transcript models than did the current direct-RNA ONT, but that the former is prone to antisense transcript artifacts that apparently stem from the reverse transcription step into cDNA. It is also likely that many of the ISM-class of transcripts that we detected more prominently in ONT are false positives due to the pore ceasing to sequence midway through an RNA. While over 80% of the transcript models we detect in the well-studied GM12878 cell line were already known, we nonetheless found evidence of a number of new NIC and NNC transcript models that are supported independently by 5' and 3' ends from other genomics assays.

In contrast to the homogeneous and frequently-studied GM12878 cell line, we found that a substantial number of genes in the mouse cortex and hippocampus produced more novel (NIC and NNC) isoform reads than known isoforms. Not surprisingly, this suggests that we are still underestimating the overall contributions of alternative splicing in tissues that are both more complex in terms of cell composition and also less comprehensively measured to date. At this time, the goal of producing a reference-level annotation transcriptome for any given cell type or tissue is well-served by the PacBio platform, but our results also make it clear that any platform that provides transcript information by direct RNA sequencing, as the RNA ONT platform now does, makes a different and important contribution. At our current PacBio sequencing depth, we do not expect to encounter substantial issues with lack of complexity in our bulk cDNA libraries. However, as long-read cDNA sequencing depths increase, reads from PCR duplicates may become much more prevalent and would be difficult to detect without UMIs. This is never an issue with direct-RNA sequencing on the Nanopore platform because each read must correspond to a distinct mRNA molecule. As iterative advances are made on these platforms, and as other long-read systems are added, the ability to process and compare the outputs from all versions of all systems in a platform-agnostic way will be increasingly important.

In addition to the technology-specific challenges of each long-read platform, we identified some shared issues. While both PacBio and ONT could sequence most genes expressed in the cells, some very long transcripts were conspicuously missing or under-represented in our data. For instance, in GM12878, we only detected 3 reads that fully matched known isoforms of the highly expressed XIST gene in terms of their splice

junctions. Even the longest of these reads (3,539 nt) was missing several kb from the 5' and 3' ends of the annotated GENCODE model. More generally, while NIC and NNC transcript models looked identical to or better than known transcripts in terms of CAGE, poly(A), and RNA-PET validation, ISMs represent a challenge for both technologies. This is particularly pressing as we detect more such ISMs in our brain tissue biosamples than in cell lines. We expect that parsing ISMs will be a challenge in human post-mortem tissue samples, including reference collection efforts for ENCODE4, because RNA quality is typically lower than what we obtained from cell lines and fresh mouse tissue sources. The "Iso-seq" approach, which intentionally enriches long-read size categories, has been to collapse ISM reads onto known transcripts. However, our results show that a subset of ISMs do have independent CAGE and 3' end support. Thus, biological ISM forms are difficult to distinguish from truncated reads without, at minimum, some independent CAGE support. Interestingly, the XIST locus is crowded with CAGE peaks throughout its longest transcript model, suggesting that there may be multiple "shorter" isoforms produced than previously appreciated, with evidence for them having been ignored due to the lack of resolution using short reads alone. ISMs are, in any case, useful models to incorporate into gene expression quantifications. With additional datasets and evidence for training, we anticipate that machine learning techniques will allow us to discriminate real ISMs from technical artifacts. Until then, it seems prudent to ignore ISMs for transcript discovery in the absence of CAGE (or similar) support.

Clear challenges remain to generate fully comprehensive, high-fidelity long-read transcriptomes because of the still relatively noisy sequencing methods and imperfectly

preserved RNAs. That said, our results show that current long-read methods are already demonstrably superior to "pooled" short-read RNA-seq for reference annotation-level transcriptomics if high quality mRNA can be extracted. The resulting gain in clarity with respect to long-range isoform structure and associated isoform-specific quantification is already substantial, although relatively high costs remain a limiting factor. At the time of this study, our long-read data costs were roughly an order of magnitude higher than their short-read counterpart, although a useful perspective is that this cost is comparable to that of short-read RNA-seq 10 years ago. We expect that long-read data will decrease similarly in cost per experiment as these platforms mature. Even in the domain of single-cell RNA-seq, which is currently thriving on short single-reads for molecule counting, long-read formats are beginning to be applied, aiming to capture the richness of isoform variation and regulation on a per-cell and per-cell-type basis[47]. That said, short-read transcriptomes will surely continue to play a prominent role for short RNA class substrates, for intractably degraded RNAs, and, increasingly, in biological settings where a few long-read transcriptomes can provide a reference against which larger numbers of companion short-read samples can be quantified. Ultimately, the transition to routine long-read transcriptome quantification will allow biologists to achieve clarity about functional mRNA isoform choices and their inferred protein products for any cell type, tissue, or disease state.

3.5 Materials and Methods

3.5.1 Sample collection and RNA extraction

GM12878 cells were grown and harvested as described in the ENCODE consortium protocols (encodeproject.org). Total RNA was extracted using the QIAGEN RNAEasy Plus kit (Cat. No. 74134). All animal experimental procedures were approved by the Institutional Animal Care and Use Committee of University of California, Irvine, and performed in accordance with the NIH Guide for the Care and Use of Laboratory Animals.

Mice were anesthetized with CO_2 and perfused with phosphate buffered saline (PBS) for 5-7 minutes until most organs are clean from blood. Hippocampus and Cortex from two 8-month male C57BL/6 mice were dissected and collected in HBSS no calcium no magnesium solution (cat. No. 14170112). Tissues were homogenized using the QIAshredder while in lysis buffer included in the QIAGEN RNAEasy Plus kit (Cat. No. 74134). Total RNA extraction was done following the vendor instructions**.** To degrade mRNA without a 5' cap, total RNA was exposed to an exonuclease treatment using Terminator™ 5´-Phosphate-Dependent Exonuclease (Cat. No. TER51020).

PacBio library preparation, sequencing, and initial data processing

Starting from the depleted RNA, we followed a modified version of the SMART-seq2 protocol to synthesize cDNA[48]. 1000 ng of cDNA were used as input for the PacBio library prep following the SMRTbell Template Prep Kit 2.0 instructions. Sequencing was done on the PacBio Sequel II machine, allocating 1 SMRT cell per biological replicate. Raw PacBio subreads were processed into circular consensus reads using the Circular Consensus step

(CCS v4.0.0) from the SMRTanalysis 8.0 software suite (parameters: --noPolish --minLength=10 --minPasses=3 –min-rq=0.9 –min-snr=2.5) **(Figure S3.1a)**. Next, adapter configurations were identified and removed using Lima v1.10.0 (parameters: **--isoseq --num-threads 12 --min-score 0 --min-end-score 0 --min-signal-increase 10 --min-score-lead 0). After this,** full-length non-chimeric (FLNC) reads were extracted using the Isoseq3 Refine step (v3.2.2; parameters: **--min-polya-length 20 --require-polya)**. This program considers a read to be FLNC if it contained the expected arrangement of 5' and 3' PacBio primers at the Lima stage as well as a poly-(A) tail. Refine orients the reads to the correct strand (reverse-complementing sequences as necessary) and removes the poly-A tails. The resulting FLNC reads were mapped to the reference genome using Minimap2 version 2.17 (GRCh38 assembly for human cell types, and mm10 for mouse) with parameters recommended by the Minimap2 documentation for PacBio (-ax splice:hq -uf --MD).

Illumina library preparation and sequencing for mouse brain samples

Starting from the same cDNA used for the mouse brain PacBio libraries, we built short-read libraries using the Nextera DNA Flex Library Prep Kit (https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-dna-flex.html?scid=2017249vu1 <%22>). These libraries were sequenced on an Illumina NextSeq500 to a minimum of 50 million paired-75bp reads per sample.

ONT library preparation and sequencing

Starting from 45 µg of depleted RNA, we proceeded to the direct-RNA library prep following the RNA-002 kit instructions. Reverse transcription was used to get rid of secondary RNA structures. We used R9.4 flowcells and MinKNOWN 2.0 was used to run the samples until having 2 millionraw reads. Basecalling was performed on the direct RNA ONT reads using ONT Guppy 3.2.1+334123b (parameters: -r --flowcell FLO-MIN106 --kit SQK-RNA002 --disable_pings -q 0 --read_batch_size 4000000 --reverse_sequence on --u_substitution on -x "cuda:0 cuda:1") **(Figure S3.1b)**. ONT reads were mapped to the reference genome using Minimap2 version 2.17. We used parameters recommended for ONT by the Minimap2 documentation (ax splice -uf -k14 -MD).

Preparing reference genomes and transcriptome annotations

The human and mouse reference genomes were obtained from the ENCODE portal (GRCh38 assembly for human cell types, and mm10 for mouse). All information other than the chromosome name was removed from the FASTA headers in the reference genome files. GENCODE v29 human and GENCODE vM21 mouse comprehensive GTF transcriptome annotations were downloaded from the GENCODE portal.

Since all samples were sequenced with ERCC spike-ins and SIRVs, it was necessary to augment the reference genomes and transcriptomes with these transcripts. The sequences of the SIRVs and ERCCs (Set 3) as well as the SIRV GTF annotation were downloaded from Lexogen here: https://www.lexogen.com/wp-content/uploads/2018/08/SIRV_Set3_Sequences_170612a-ZIP.zip. To create augmented

reference genomes, we concatenated each of the human and mouse fasta files with SIRV_ERCCs_multi-fasta_170612a.fasta and SIRV_isoforms_multi-fasta_170612a.fasta.

Additional processing was needed before adding the SIRV and ERCC transcripts to the human and mouse annotations. Since no GTF file was provided for the ERCCs, we created one by running *merge_encode_annotations.py* on the ERCC fasta file. The SIRV isoforms (SIRV_isoforms_multi-fasta-annotation_C_170612a.gtf) were processed with the *separate_multistrand_genes.py* script so as to separate transcripts located on different strands into separate genes. Next, we ran *talon_reformat_gtf* (TALON utility) on the ERCC and SIRV GTFs in order to add in explicit gene and transcript lines needed by the TALON program. These reformatted GTF files were then concatenated to the end of the human and mouse GENCODE annotations.

TALON pipeline

Following alignment to the genome, reference-based error correction was performed on the PacBio FLNC and ONT reads using TranscriptClean v2.0.2 (available on GitHub at https://github.com/mortazavilab/TranscriptClean). Reference splice junctions were derived from the GENCODE annotations using TranscriptClean accessory script *get_SJs_from_gtf.py*. For the human runs, we used VCF-formatted NA12878 truth-set small variants from Illumina Platinum Genomes to run TranscriptClean in variant-aware mode (--canonOnly + defaults). For the mouse datasets, we ran TranscriptClean without a VCF file (--canonOnly + defaults). By using the –canonOnly flag, we omitted any reads that still contained one or more un-annotated noncanonical splice junctions from the output.

After TranscriptClean, we ran the TALON module **talon_label_reads** on each corrected SAM file in order to compute the fraction of As following the end of each read alignment. We set the –ar parameter to 20 bp to match the length of the T sequence used in PacBio's oligo-dT primer for poly-(A) capture. All TALON steps (including this one) were run with version 5.0. TALON and accompanying documentation are available from https://github.com/mortazavilab/TALON.

Human and mouse TALON databases were initialized from the GENCODE v29 and GENCODE vM21 + SIRVs/ ERCC annotations using the **talon_initialize_database** module from the TALON package (parameters: --l 0 --5p 500 --3p 300). To annotate the GM12878 PacBio and ONT reads, we created a configuration file with all four datasets in it and ran the **talon** module on this file along with the human TALON database (parameters: --cov 0.9 --identity 0.8). To annotate the mouse cortex and hippocampus reads, we created a configuration file with all four datasets in it and ran the **talon** module on this file along with the mouse TALON database (parameters: --cov 0.9 --identity 0.8).

To perform long read quantification, transcript abundance matrices were extracted from the TALON databases using the **talon_abundance** module. We used the unfiltered abundance files for all gene-level expression analyses (omitting genomic transcripts). To perform transcript-level analyses, we first used the **talon_filter_transcripts** utility to generate celltype and experiment-specific transcript whitelists (parameters: --maxFracA 0.5 –minCount 5). This filtering process selected for transcript models that were 1) known in GENCODE/SIRV/ERCC, or 2) reproducibly detected at least 5 times in each specified dataset. Reads with > 0.5 fraction As (as specified by **talon_label_reads**) were omitted when

computing this read support. We generated separate whitelists for the PacBio GM12878, ONT GM12878, PacBio cortex, and PacBio hippocampus dataset pairs. The resulting whitelists were used to generate filtered abundance files for transcript quantification (using **talon_abundance)**, as well as custom filtered GTF annotations (using **talon_create_GTF***)*.

Further details and custom scripts for data visualization are available on GitHub (https://github.com/mortazavilab/TALON-paper-2020).

PacBio vs. Illumina short read comparison

Illumina short-read RNA-seq reads from GM12878 were downloaded from the ENCODE portal in the fastq format (accession ENCSR000AEH). Quantification against the GENCODE v29 annotation was performed on each biological replicate using Kallisto[16]. The log fold changes between PacBio and Illumina counts for each GENCODE gene/transcript were computed using the exact test method in EdgeR (v3.28.1) following filtering of lowly expressed genes/transcripts and normalization. Bonferroni correction for multiple testing was performed on the resulting p-values. Genes/transcripts were considered significantly different in the two platforms if adjusted $p < 0.01$ and abs(log2FC) > 1.

In addition, we computed the gene-level Spearman correlations between each long-read technology and Illumina for GM12878 for genes that were detected by both platforms. To do this, we first averaged the expression (in TPM) of each gene across biological replicates by platform. For Illumina, this meant averaging the Kallisto TPM results across replicates. For PacBio and ONT, we computed the gene-level TPMs from the unfiltered TALON abundance tables for each dataset (excluding genomic transcripts and novel genes), then averaged the replicates.

Comparison of PacBio and ONT transcriptomes

We calculated gene quantification using the unfiltered TALON abundance files with genomic transcripts removed. For transcript quantification, we included transcript models in the union of the PacBio and ONT filtering whitelists.

CAGE analysis

Robust human CAGE peaks were downloaded from FANTOM5 in the BED format[12]. The genomic coordinates were mapped from hg19 to hg38 using the UCSC genome browser LiftOver tool[49]. We obtained the start site of each long-read transcript model from our GTF transcriptomes, then used Bedtools to ascertain whether any CAGE peak overlapped the 100 bp region immediately up or downstream of each TSS[50].

Computational Poly(A) motif analysis

Each GTF transcript model was converted to BED format. We extracted the DNA sequence of the last 35 bp in each transcript using the reference genome (GRCh38 assembly for human cell types, and mm10 for mouse), then searched for the presence of a known 6-mer poly(A) motif as described in Anvar *et al.*, 2018[51].

RNA-PET analysis

RNA-PET clusters for GM12878 were downloaded in the BED format from the ENCODE portal (accession ENCFF001TIL). The genomic coordinates were mapped from hg19 to hg38 using the UCSC genome browser LiftOver tool[49]. We obtained the start and end

site of each long-read transcript model from our GTF transcriptomes, then used Bedtools to check whether any pair of RNA-PET clusters was located within 100 bp of the start and end[50].

Mouse Hippocampus and Cortex data analysis

Gene and transcript abundances were calculated as described above. For differential transcript expression analysis, we used EdgeR (v3.28.1) and adjusted the resulting p-values using the Bonferroni method. Transcripts with abs(log2FC) > 1 and an adjusted p-value < 0.01 were considered significantly differentially expressed. We used a custom script to identify genes that had higher novelty counts (NIC+NNC) separately for cortex and hippocampus and identified the overlapping genes. The UCSC genome browser was used to visualize transcripts colored according to their novelty.

Overview of TALON database

A key novel aspect of the TALON pipeline is its use of a database to store transcript models and abundances from multiple runs, and therefore its ability to compare new datasets to this knowledge base. The database is designed to serve two major purposes:

1) To store gene, transcript, and exon attributes of the type needed to construct a GTF transcriptome annotation (i.e. their names/IDs, genomic positions, and novelty type).

2) To track the quantity and identity of the transcripts observed in each of the datasets that have been processed so far.

The underlying philosophy of TALON filtering is that as additional datasets are sequenced and added to the database, more information becomes available to differentiate

between real transcripts and artifacts. Therefore, it makes sense to apply filtering to novel transcript models in the database as a post-processing step that can be revisited at any time, rather than discarding transcripts upfront during a run. Datasets can be processed back-to-back with TALON as part of a single run but can also be added successively without the need to re-analyze the earlier data since the results are already tracked in the database.

TALON database structure

Each instance of a TALON database consists of 14 tables in total in the SQLite format **(Figure S3.10)**. The database is initialized from a GTF-formatted transcriptome annotation such as GENCODE, which populates its 'gene_annotations', 'transcript_annotations', and 'exon_annotations' tables with the metadata from each GTF entry. Notably, these tables permit data to be entered from more than one source, recognizing for example that it is possible for a transcript to have a different name or novelty status depending on the particular annotation version consulted.

During initialization, the locations of the genes, transcripts, and exons must also be stored. Rather than placing genomic coordinates directly in the 'gene' or' transcript' tables, we considered how the database could be extended in the future to accommodate personalized genomes for human transcriptome analysis, or genomes of different mouse strains. Individual genomic coordinates are abstracted out and represented by a vertex ('vertex' table), which can have a different location depending on the instance of the genome build in the database (as denoted in the 'location' table). Exons and introns are represented as edges connecting two vertices, which means that transcripts are paths through vertices

belonging to a gene. These are stored in the 'edge' table. In the future, this graph structure could be exploited by superimposing count data onto it and examining the probability of different transcripts.

The database also contains two major tables for the explicit purpose of tracking transcripts in long read datasets. The 'abundance' table stores the number of times each transcript was detected in each dataset, which is highly useful for quantitative comparisons. The 'observed' table contains a record of every long read processed by the annotation pipeline. It tracks the read length, the transcript and gene assignment of the read, and any differences from the annotation at the 5' and 3' ends. The latter is important because more accurate 5' and 3' ends are a major goal of long read transcriptomic analysis. The 'dataset' table tracks associated metadata for each dataset that was initially entered by the user in the TALON configuration file.

Epstein-Barr Virus transcriptome analysis

An EBV chromosome GTF annotation was obtained from https://ebv.wistar.upenn.edu/downloadstatis/ebv.custom and refined for use. PacBio GM12878 reads that mapped to the EBV chromosome from the hg38 genome build were isolated and run through TranscriptClean using splice junctions generated from the GTF, and subsequently run through TALON. Gene and transcript TPMs were calculated using previously discussed filtering methodologies.

Running FLAIR on PacBio and ONT data

FLAIR was cloned from BrooksLabUCSC/flair on GitHub on 3/12/20 (latest commit d23a9c2ef62ede402e8b23d6231784ad910ed1af). For each of the PacBio and ONT GM12878 datasets, we ran the FLAIR align and correct steps on biological replicates separately, then combined the outputs in order to run the FLAIR collapse and quantify steps using default parameters. We removed SIRV and ERCC transcripts at this point and converted the 'counts_matrix.tsv' output file from FLAIR Quantify to a format resembling the TALON abundance file. From there, we compared gene and transcript detection to the TALON results and to Illumina.

Long-read splice junction extraction

Post-TALON splice junctions and GENCODE annotation splice junctions were extracted from GTF files using the get_SJs_from_gtf.py script from TranscriptClean (v2.0.2).

Short-read splice junction extraction

To obtain high-confidence splice junctions from short reads, Illumina RNA-seq reads (fastq) were mapped to the reference genome using STAR v. 2.5.2a. We used the following ENCODE-recommended parameters:

```
STAR \
--runThreadN 4 \
--genomeDir genome \
 --readFilesIn illumina_1.fastq illumina_2.fastq \
--sjdbGTFfile gencode.annotation.gtf \
--outFilterType BySJout \
--outFilterMultimapNmax 20 \
--alignSJoverhangMin 8 \
```

```
--alignSJDBoverhangMin 1 \
--outFilterMismatchNmax 999 \
--outFilterMismatchNoverLmax 0.04 \
--alignIntronMin 20 \
--alignIntronMax 1000000 \
--alignMatesGapMax 1000000 \
--outSAMattributes NH HI NM MD jM jI \
--outSAMtype SAM
```

For each splice junction, the resulting file lists genomic location, strand, intron motif, whether or not the junction is annotated in GENCODE, and the amount of read support. To ensure splice junction reproducibility, we ran each replicate separately, and the subsequent splice junctions were merged and filtered for each cell type. Splice junctions with no supporting uniquely-mapped reads were discarded, and we required non-annotated splice junctions to have at least one uniquely-mapping read in each replicate.

Splice junction support by junction novelty category

To determine the novelty of all long-read, post-TALON GM12878 splice, we first extracted each junction from the TALON GTF. Next, we used a custom script to compare each long-read junction with the splice donors and acceptors present in the GENCODE annotation. We defined three different junction categories:

- Known junctions: The exact splice donor/acceptor combination was seen in the GENCODE annotation.
- Novel in catalog: The splice donor and acceptor are seen in the GENCODE annotation, but never together in the same junction.
- Novel not in catalog: The splice donor, splice acceptor, or both fail to be seen in the GENCODE annotation.

Once the status of each junction was determined, we computed the GM12878 short-read support for each splice junction novelty category separately by platform.

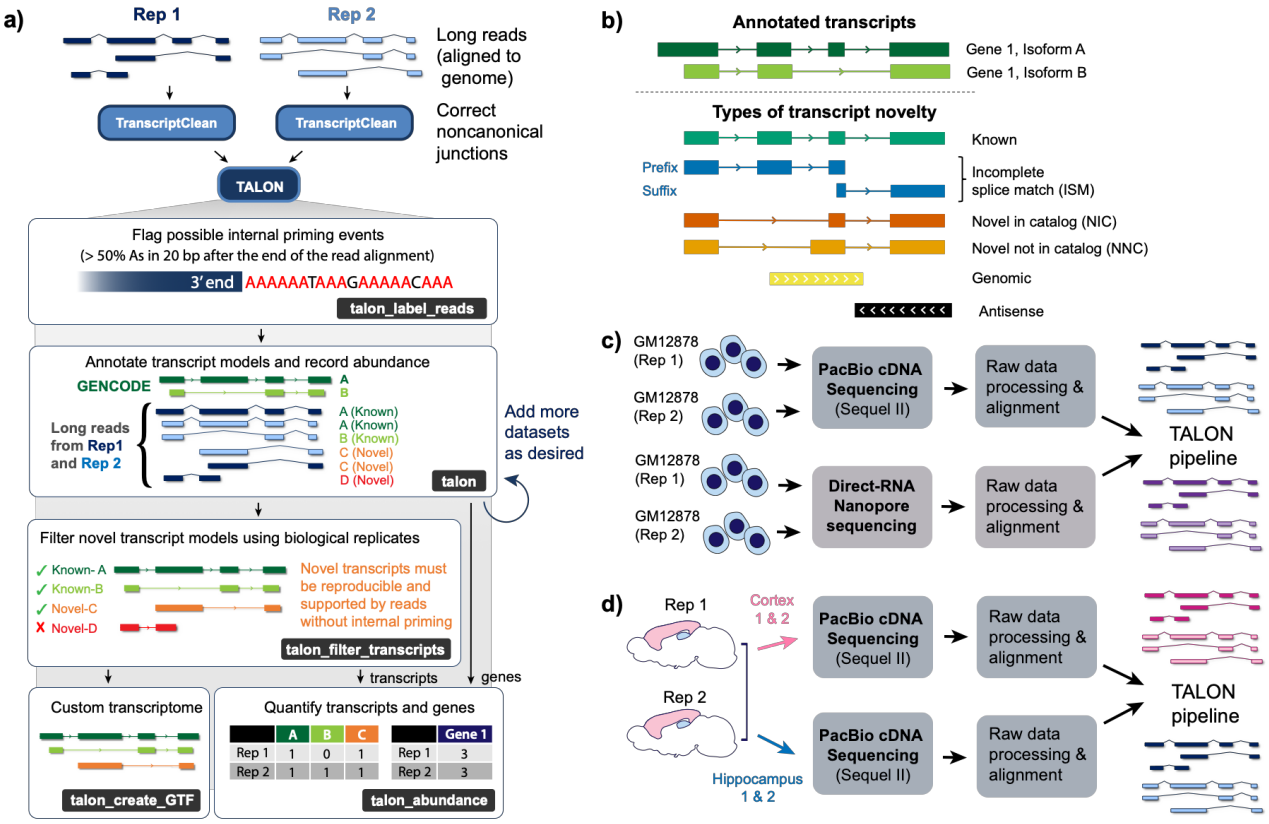

Figure 3.1. Overview of TALON. a) Long-read alignments from either technology are corrected with TranscriptClean for each biological replicate. Next, potential internal priming events are flagged by the talon_label_reads module. Labeled reads are passed into talon, where they are assigned a gene and transcript identity. The talon_abundance module computes gene expression directly from the talon results, whereas novel transcript models are filtered prior to quantification. Novel transcripts must be reproducibly detected n times in k datasets to pass the filter (default n = 5 and k = 2), and must not come from internally primed reads. **b)** Types of transcript novelty tracked by TALON. **c)** TALON can be used to compare different long-read sequencing technologies run on the same biological sample such as the human GM12878 cell line. **d)** TALON can also be used to compare genes and transcripts across different samples such as mouse hippocampus and cortex.

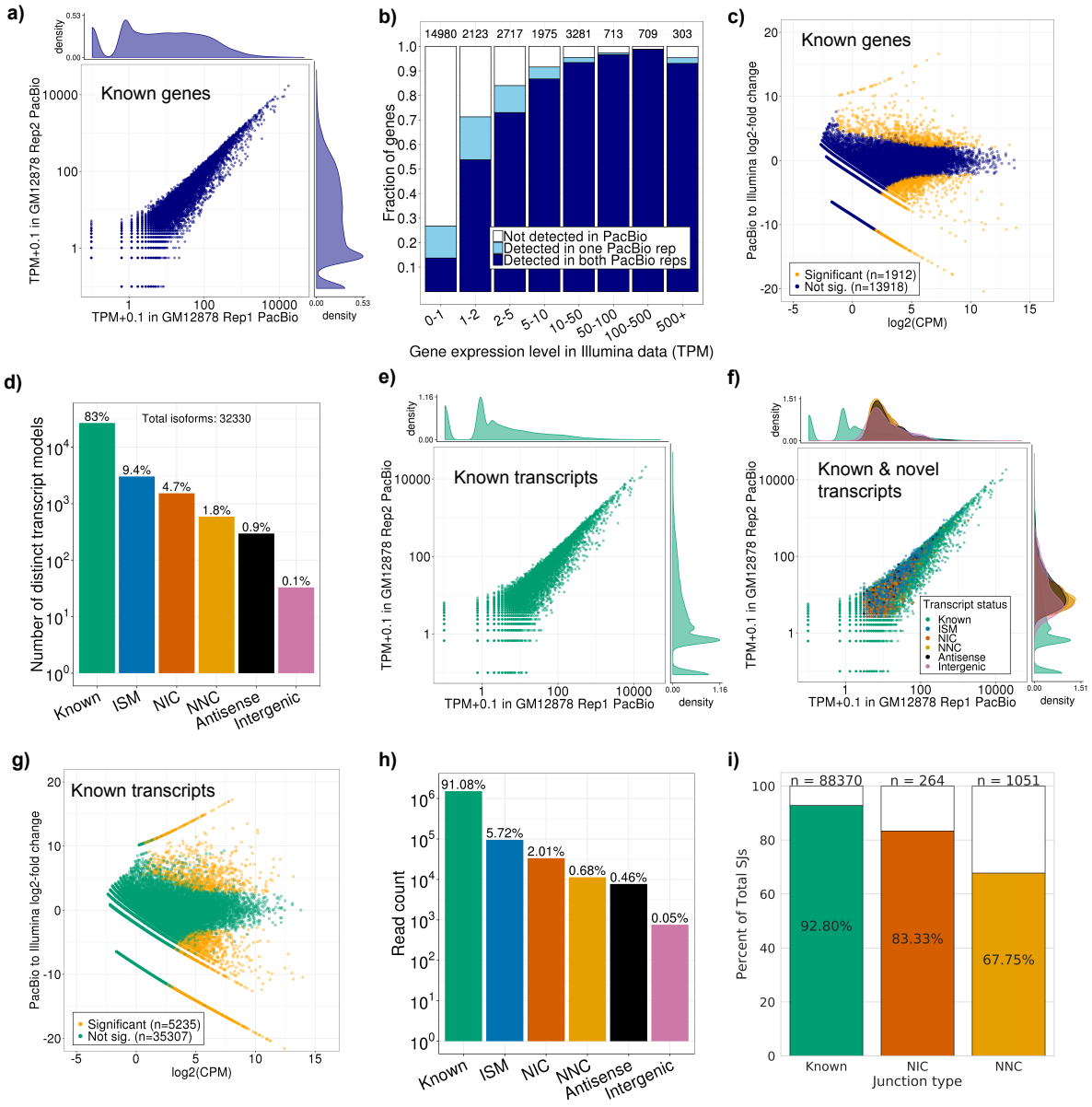

Figure 3.2. Performance of TALON on PacBio transcripts from GM12878 cell line. a) Expression level of known genes (GENCODE v29) in each biological replicate (Pearson r = 0.97, Spearman rho = 0.92). **b)** Proportion of genes expressed in Illumina RNA-seq data of GM12878 that are also detected in PacBio, binned by Illumina expression level (TPM). **c)** Comparison of gene expression levels for known genes in the PacBio and Illumina RNA-seq platforms. **d)** Number of distinct transcript isoforms observed in each novelty category. **e)** Expression level of known transcript models in each biological replicate (Pearson r = 0.97, Spearman rho = 0.73). **f)** Expression of transcript models in each biological replicate, labeled by their novelty assignments (Pearson r = 0.97, Spearman rho = 0.83) . **g)** Comparison of known transcript expression levels in the PacBio and Illumina RNA-seq platforms. **h)** Total number of PacBio reads assigned to each novelty category after transcript filtering (Rep 1). **i)** Percentage of known and novel PacBio GM12878 splice junctions supported by Illumina. Junctions labeled NIC indicate novel combinations of known splice sites, while NNC junctions included a new donor and/or acceptor.

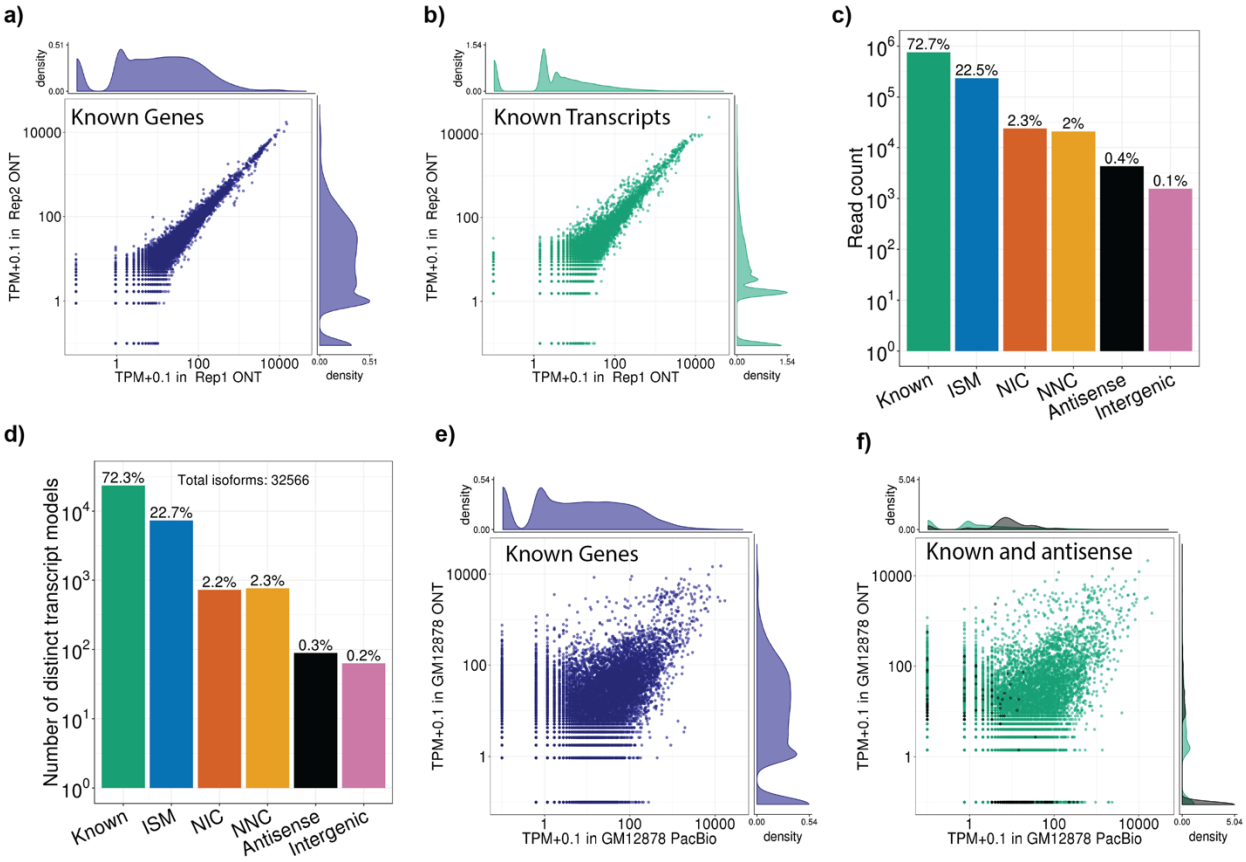

Figure 3.3. Comparison of Oxford Nanopore direct RNA-seq transcriptome with PacBio transcriptome in GM12878. a-b) 2 GM12878 replicates were sequenced using the MinIon platform and analyzed using TALON pipeline to measure **a)** gene expression (Pearson r = 0.99, Spearman rho = 0.92) and **b)** transcript expression (Pearson r = 0.97, Spearman rho = 0.64). **c)** Total read count per novelty category. There is a substantially larger fraction of ISM reads than full-length known compared to PacBio (Fig 2h). **d)** Number of distinct isoforms by novelty category. **e-f)** Correlations between ONT direct RNA-seq and PacBio with respect to **e)** gene expression (Pearson r = 0.58, Spearman rho = 0.63) and **f)** transcript expression (Pearson r = 0.5, Spearman rho = 0.18).

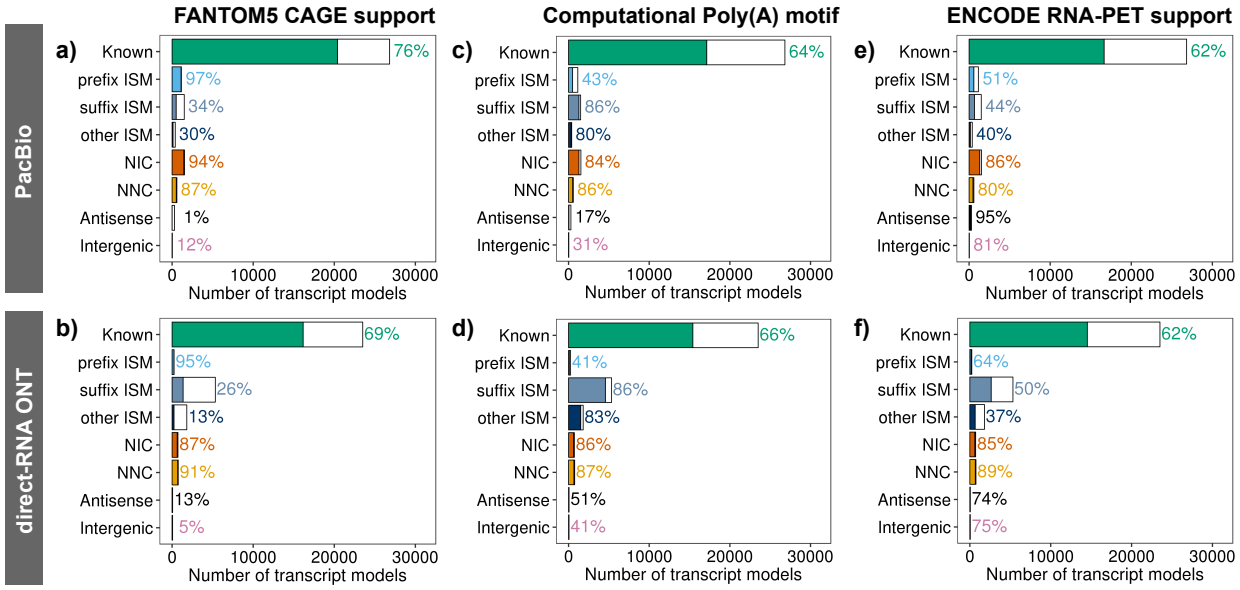

Figure 3.4. External validation of transcript model ends by novelty category. a) Percentage of TALON transcript models with CAGE support for their 5' end by novelty category (GM12878 PacBio). **b** Percentage of TALON transcript models with a poly(A) motif identified at their 3' end (GM12878 PacBio). **c)** Percentage of TALON transcript models with RNA-PET support for their 5'-3' end pair (GM12878 PacBio). **d)** Percentage of TALON transcript models with CAGE support for their 5' end by novelty category (GM12878 ONT). **e)** Percentage of TALON transcript models with a poly(A) motif identified at their 3' end (GM12878 ONT). **f)** Percentage of TALON transcript models with RNA-PET support for their 5'-3' end pair (GM12878 ONT).

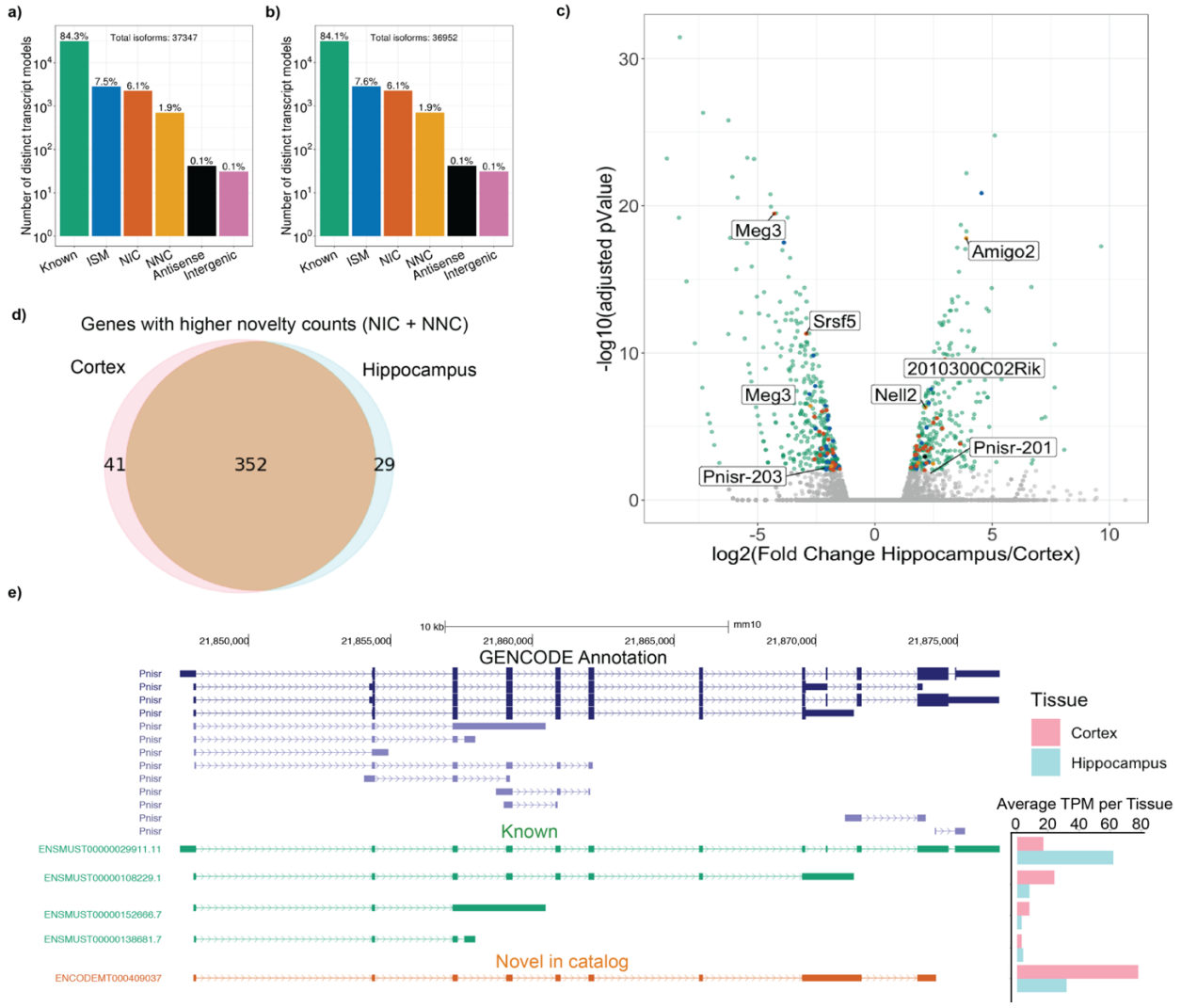

Figure 3.5. PacBio transcriptomes of 8-month male adult mouse cortex and hippocampus. Novelty assignments of distinct transcript models detected in one representative replicate each of **a)** cortex and **b)** hippocampus. **c)** Differential isoform expression in hippocampus and cortex. Transcripts with a fold change > 1 and an adjusted p value of < 0.01 are colored according to their novelty status and labeled with their corresponding gene name (color scheme in panels a,b) **d)** Detection of genes with greater novel read counts (NIC + NNC) than known. **e)** Pnisr UCSC genome browser visualization showing a new combination of exons detected by PacBio. Expression levels of each isoform detected are is plotted on the right for cortex and hippocampus.

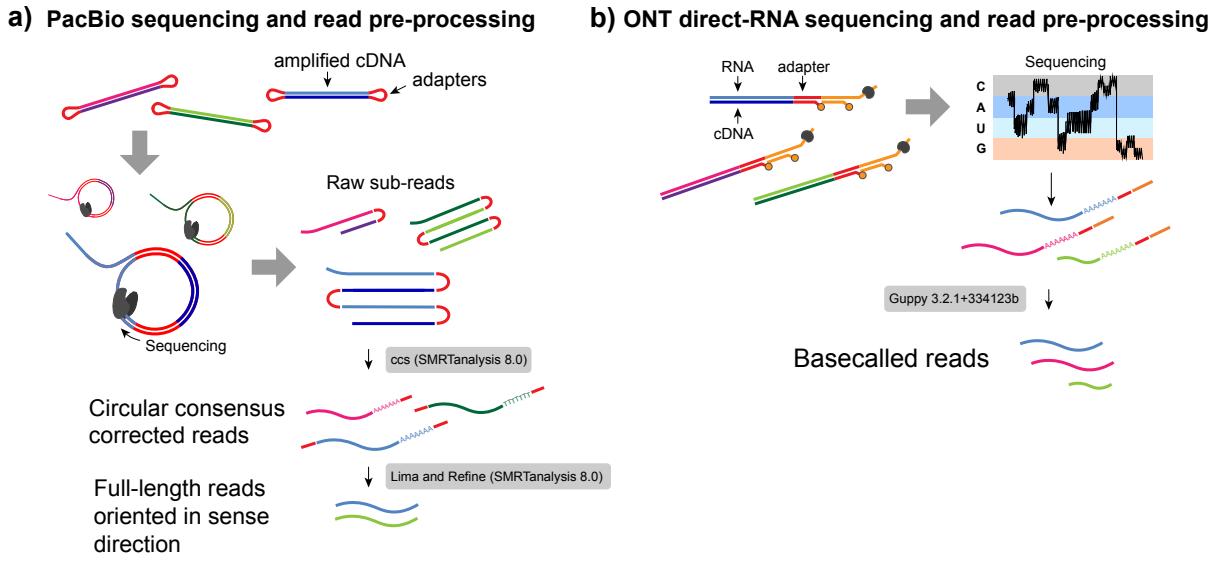

Figure S3.1: Platform-specific data processing performed prior to running TALON pipeline. a) Sequencing and preprocessing of PacBio Sequel data. The Lima/Refine step in particular is important because it removes reads that did not receive a full sequencing pass and orients the remaining reads to the correct strand. **b)** Sequencing and preprocessing of ONT direct-RNA data. Since the RNA itself is sequenced poly(A) first, no additional read orientation steps are required.

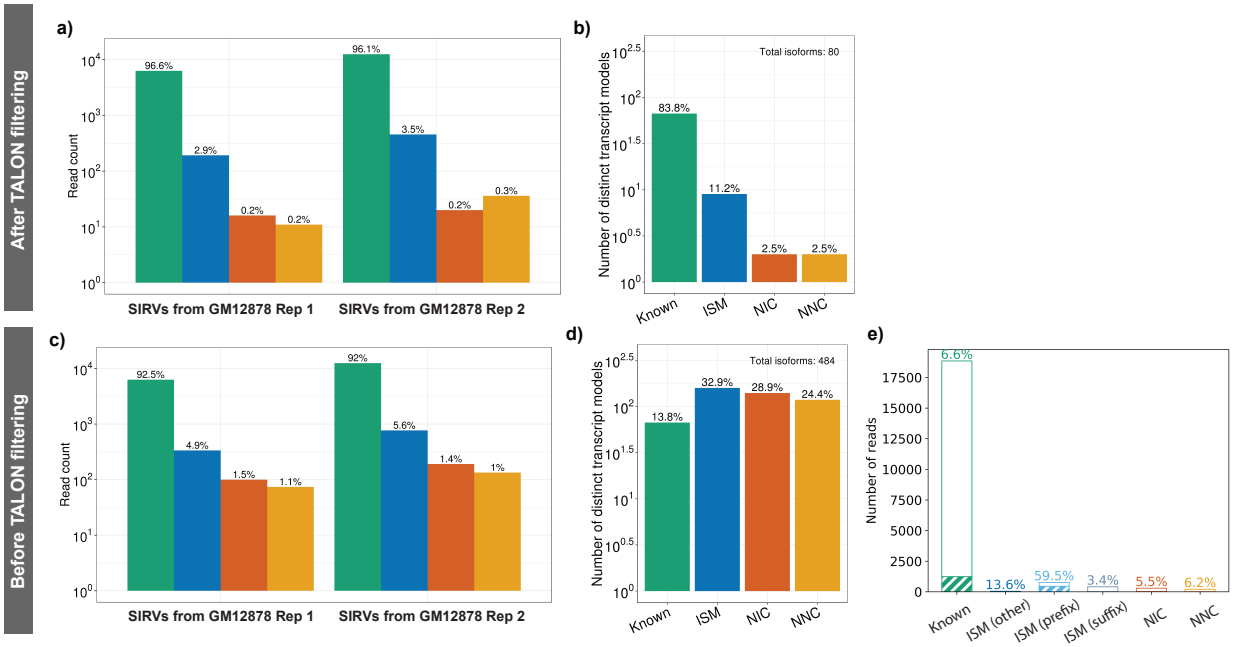

Figure S3.2. Performance of TALON filtering on SIRV transcripts sequenced with PacBio Sequel II. a) Number of SIRV-aligned reads assigned to each transcript novelty category in the GM12878 Rep1 and Rep2 datasets after TALON filtering. **b)** Number of distinct transcript models called per novelty category from the SIRV-aligned reads after TALON filtering. Union of GM12878 Rep1 and Rep2 is shown. **c)** Number of SIRV-aligned reads assigned to each transcript novelty category in the GM12878 Rep1 and Rep2 datasets (no filtering). **d)** Number of distinct transcript models called per novelty category from the SIRV-aligned reads (no filtering). Union of GM12878 Rep1 and Rep2 is shown. **e)** Proportion of unfiltered SIRV reads in each novelty category that display evidence of internal priming (> 50% As in 20bp window following the alignment). Union of GM12878 Rep1 and Rep2 is shown.

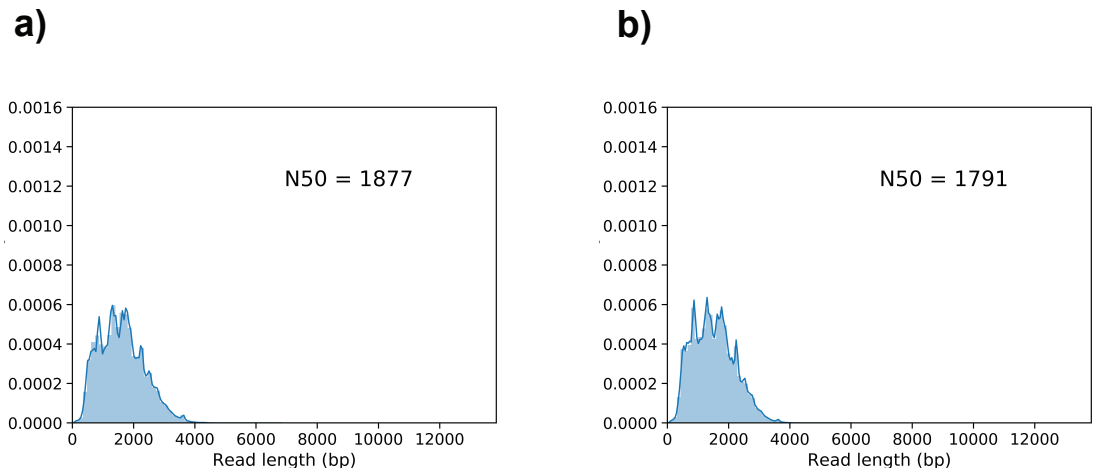

Figure S3.3. TALON read length distributions for PacBio GM12878 datasets. a) Rep 1. **b)** Rep 2.

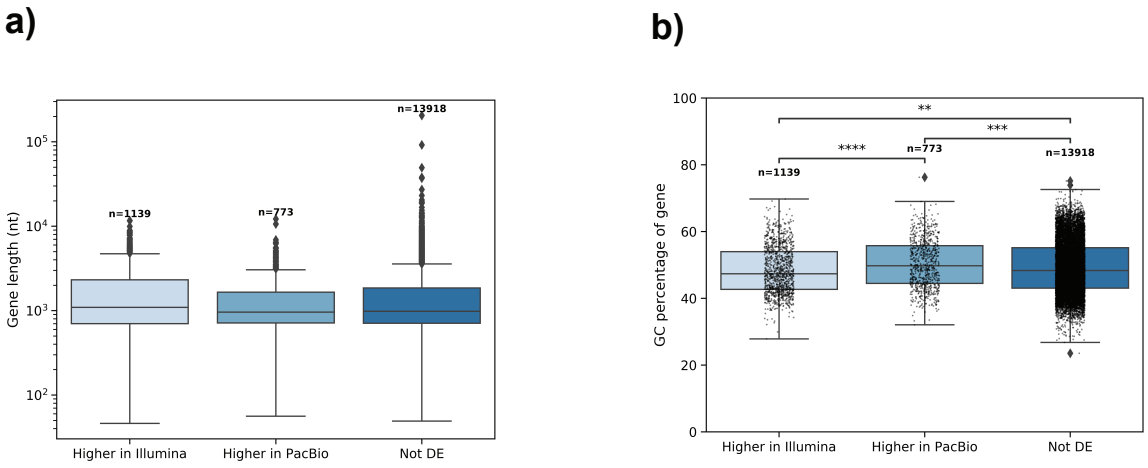

Figure S3.4. Further characterization of gene detection in GM12878 by short reads and PacBio long reads. a) Length of known genes by differential expression category. Gene length was computed by taking the median length of all known transcripts per gene. **b)** GC content of known genes by differential expression category. Gene GC content was computed by taking the median GC of all known transcripts per gene.

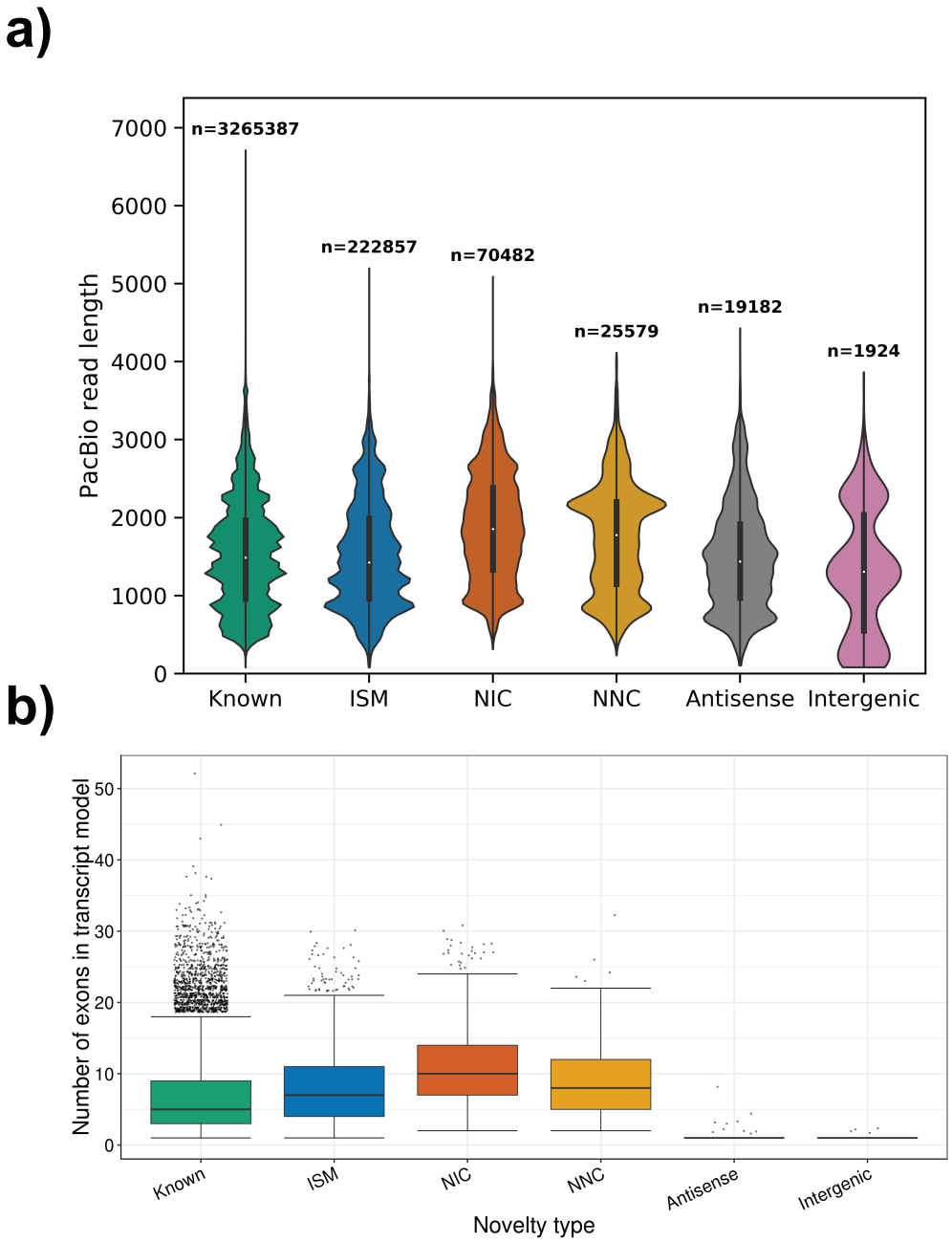

Figure S3.5. Length and exon count by transcript novelty type in GM12878 PacBio. a) Read length distributions by novelty category. **b)** Number of exons per transcript model, grouped by novelty type assignment**.**

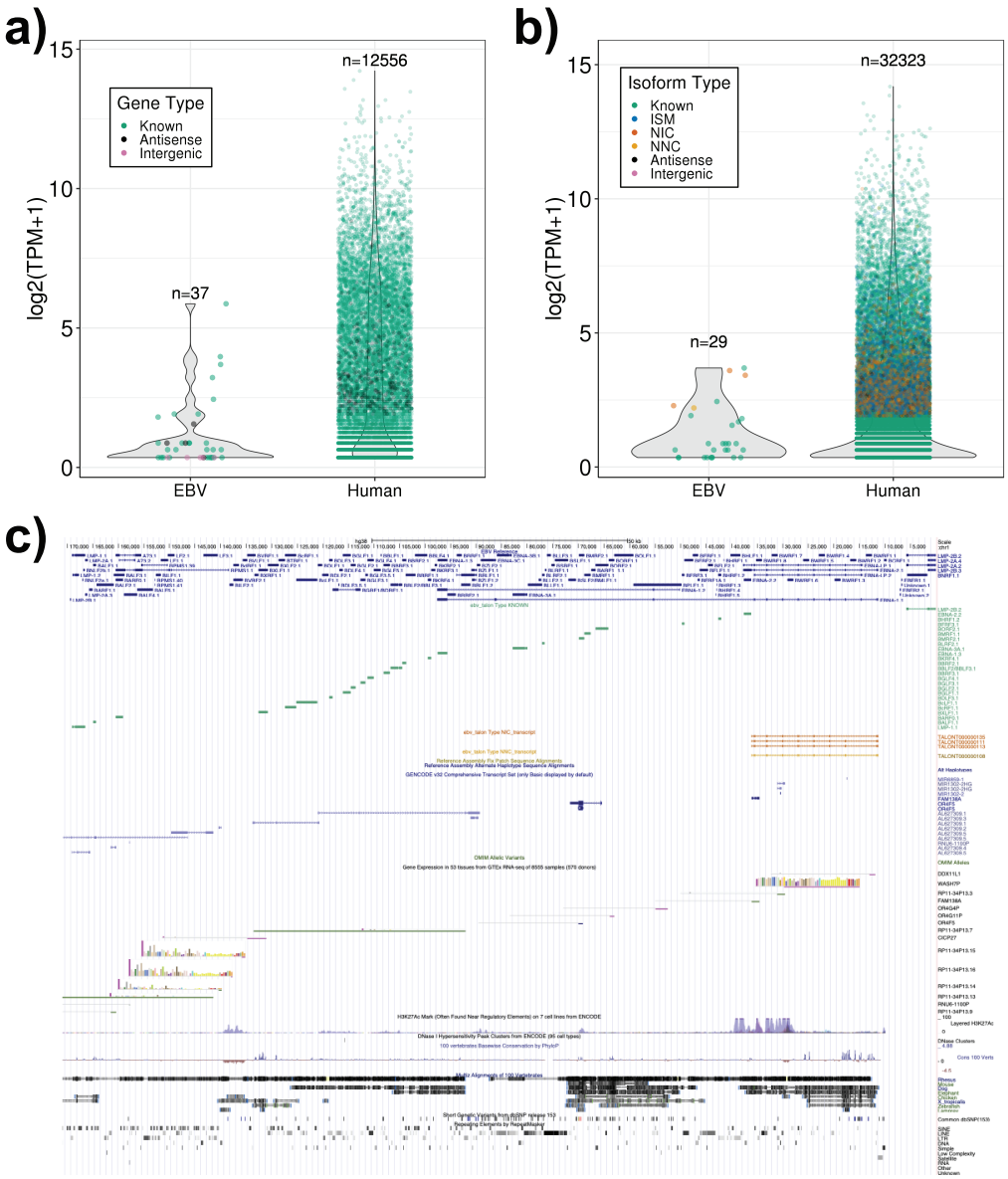

Figure S3.6. Epstein-Barr Virus transcriptome characterization in GM12878. a) Gene expression levels in GM12878 from the EBV chromosome and from the human chromosomes, labelled by gene novelty. **b)** Transcript expression levels in GM12878 from the EBV chromosome and from the human chromosomes, labelled by transcript novelty. Novel transcripts have been filtered for reproducibility between GM12878 biological replicates. **c)** Visualization of TALON GTF annotations in the UCSC genome browser for EBV transcripts in GM12878.

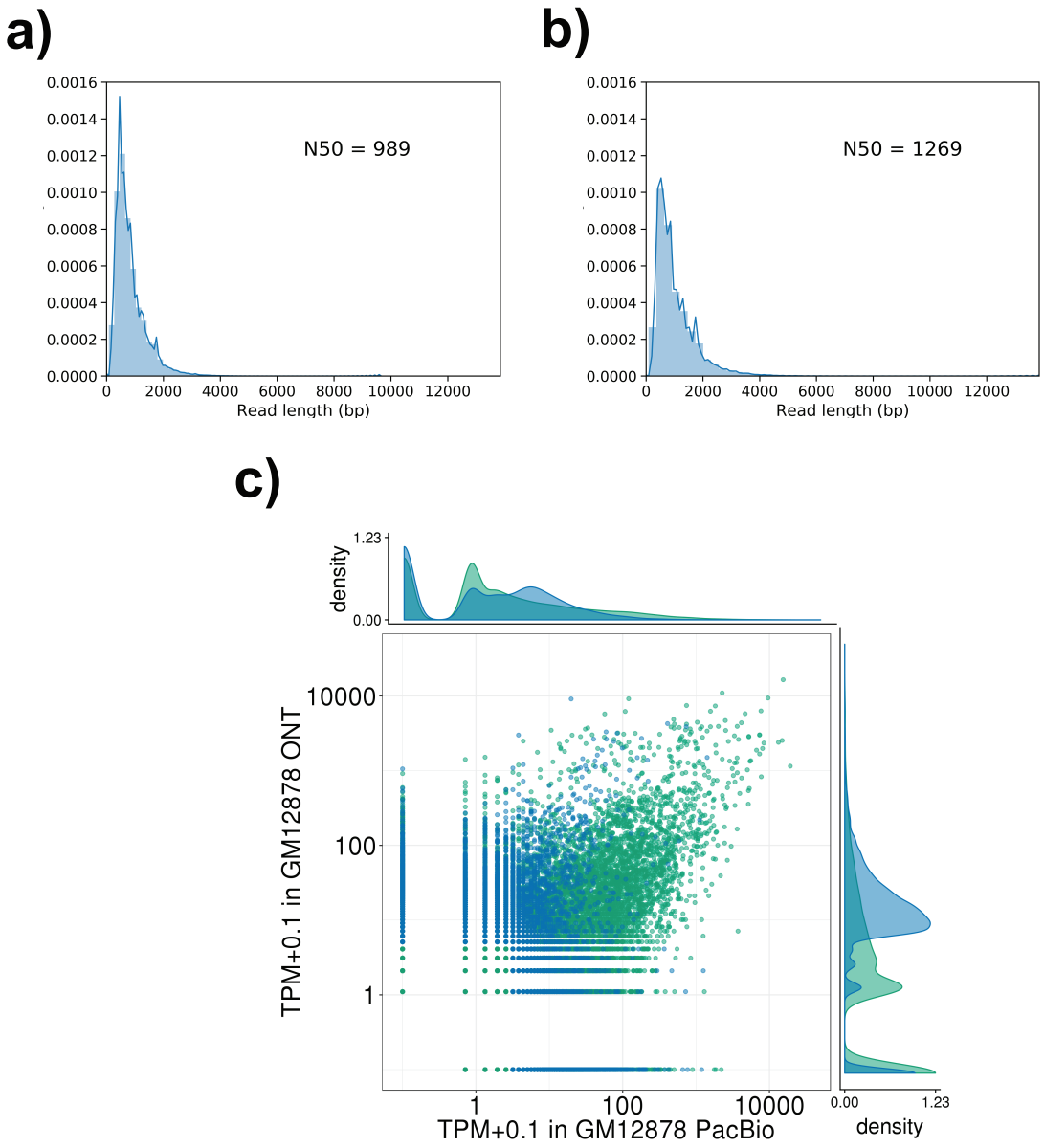

Figure S3.7. Characterization of GM12878 cell line by Oxford Nanopore direct-RNA sequencing. TALON read length distributions for Nanopore ENCODE Tier 1 cell line datasets **a)** GM12878 Rep 1 and **b)** GM12878 Rep 2. **c)** Expression level of known transcript models and reproducible ISMs in PacBio vs. ONT for GM12878 (Pearson r = 0.48, Spearman rho = 0.08).

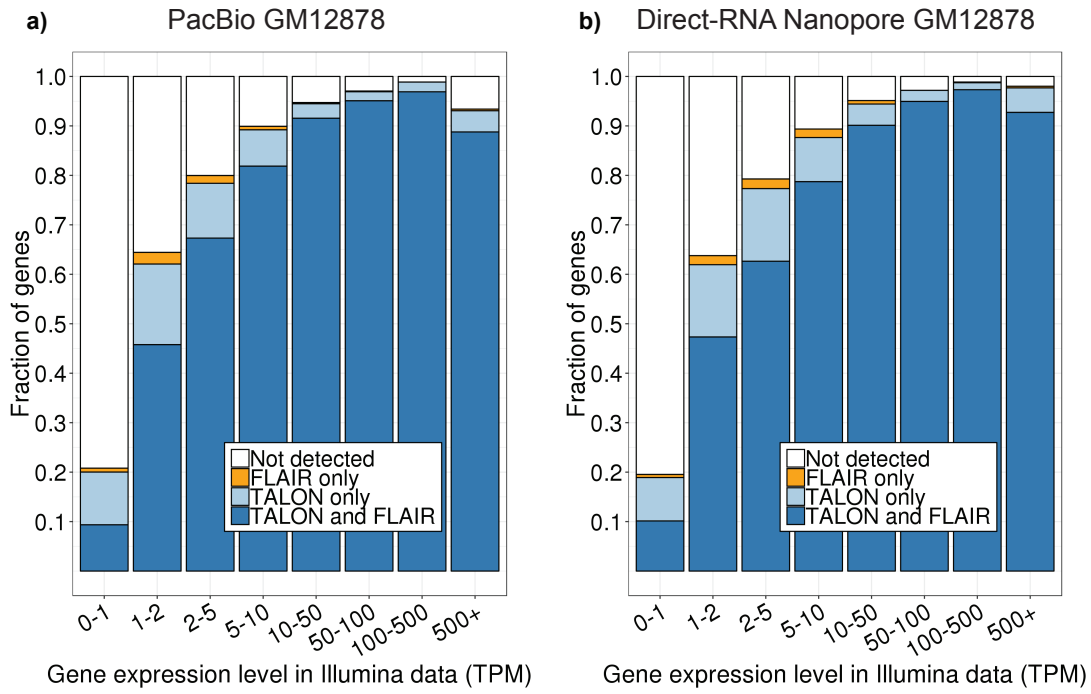

Figure S3.8. TALON and FLAIR gene detection across sequencing platforms and samples. Proportion of genes expressed in Illumina GM12878 RNA-seq data that are also detected by TALON, FLAIR, or both in the corresponding a) PacBio and **b)** ONT long-read datasets. Genes are divided into bins based on their Illumina expression level (TPM).

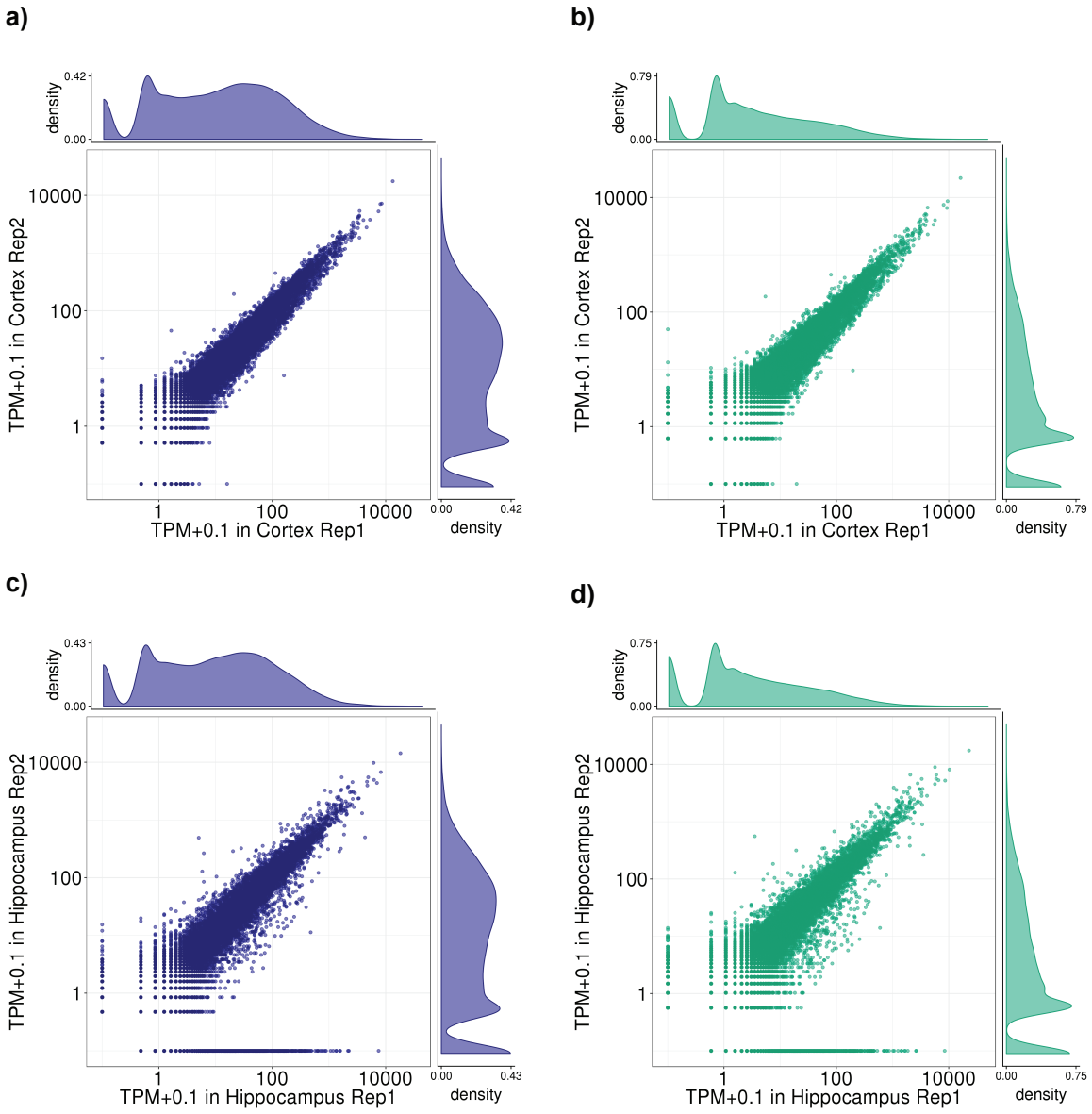

Figure S3.9. Reproducibility of PacBio gene and transcript expression in mouse cortex and hippocampus. a) Expression level of known genes in each cortex biological replicate. **b)** Expression level of known transcripts in each cortex biological replicate. **c)** Expression level of known genes in each hippocampus biological replicate. **d)** Expression level of known transcripts in each hippocampus biological replicate.

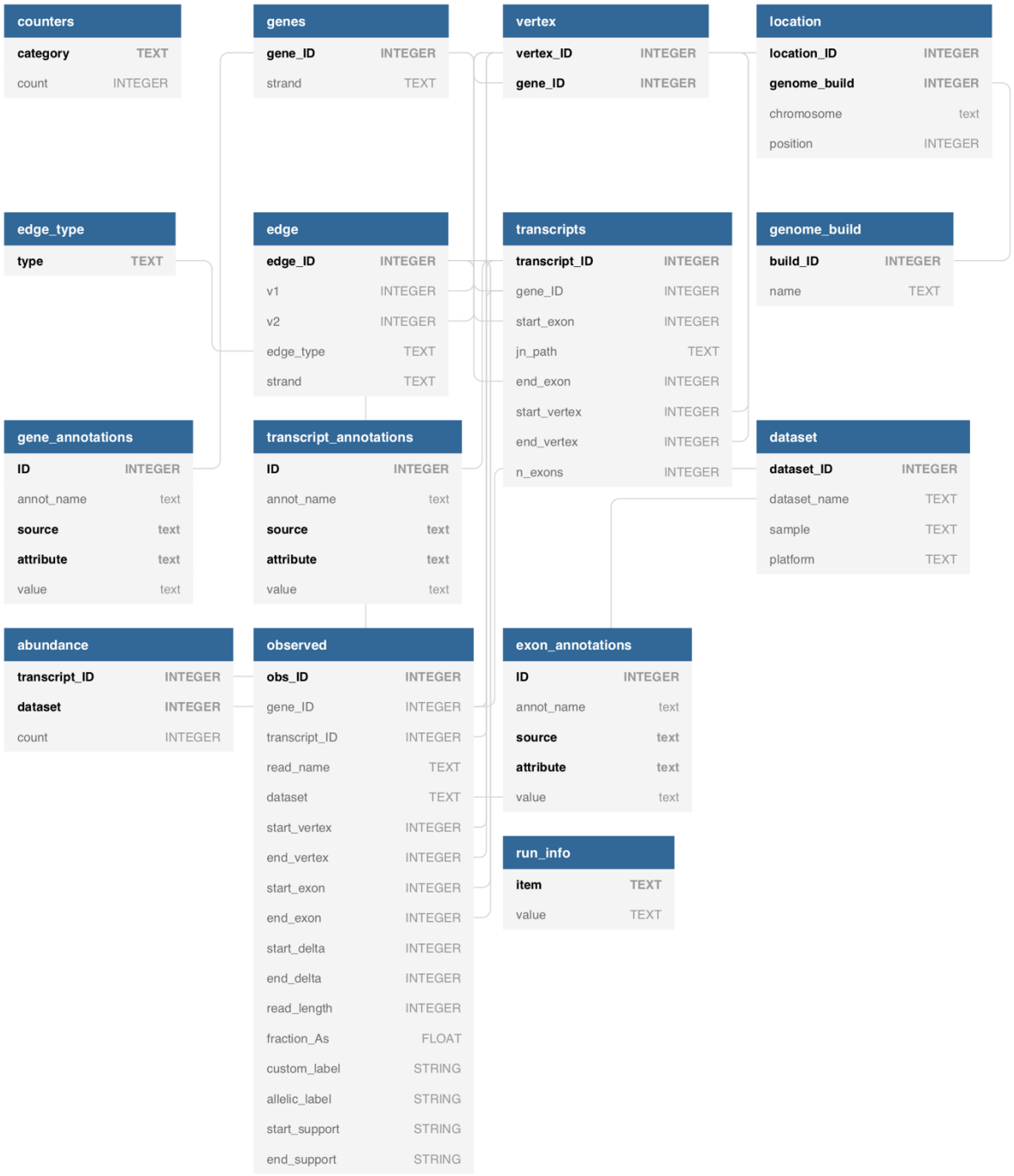

Figure S3.10. TALON database schema. Relationships between the 14 tables are indicated with grey lines, and primary keys are shown in bold.

Table S3.1: Accessions for submitted data

Platform	Cell/tissue type	Replicate	ENCODE accession	GEO accession	Raw read count	Pre-mapping read count	Reads at TALON stage
PacBio Sequel II	GM12878	1	ENCLB200YVA	---	6,061,818	2,137,168	2,040,933
PacBio Sequel II	GM12878	2	ENCLB735WVC	---	6,692,215	2,538,701	2,445,556
PacBio Sequel II	Mouse cortex	1	ENCLB287KUK	---	6,404,493	2,843,245	2,777,090
PacBio Sequel	Mouse cortex	2	ENCLB440QNX	---	6,549,444	2,643,160	2,578,722
PacBio Sequel	Mouse hippocampus	1	ENCLB722NJT	---	7,422,892	2,961,269	2,900,630
PacBio Sequel	Mouse hippocampus	2	ENCLB186LWF	---	6,943,825	3,124,583	2,858,450
ONT direct-RNA	GM12878	1	---	GSM4417547	2,020,127	2,020,127	1,675,608
ONT direct-RNA	GM12878	2	---	GSM4417548	2,571,101	2,571,101	1,984,953
Illumina RNA-seq	Mouse cortex	1	ENCLB894RIO	---	87,966,793	NA	NA
Illumina RNA-seq	Mouse cortex	2	ENCLB671GZH	---	51,152,278	NA	NA
Illumina RNA-seq	Mouse hippocampus	1	ENCLB591DUT	---	61,562,264	NA	NA
Illumina RNA-seq	Mouse hippocampus	2	ENCLB626JBH	---	62,561,081	NA	NA

Table S3.2: Detection of Illumina-expressed genes by TALON and FLAIR in PacBio GM12878

GENCODE transcripts	PacBio GM12878	ONT GM12878
Detected by FLAIR only	471	923
Detected by TALON only	12,741	11,642
Detected by both TALON and FLAIR	14,100	11,891

Table S3.3: Detection of known transcripts by TALON and FLAIR in PacBio GM12878

Illumina-expressed Genes	PacBio GM12878	ONT GM12878
Not detected	13,583	13,788
Detected by FLAIR only	234	246
Detected by TALON only	2,525	2,381
Detected by both TALON and FLAIR	10,459	10,386

3.6 References

1. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).

2. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**, 805–17 (2016).

3. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12**, 715–729 (2011).

4. Foulkes, N. S. & Sassone-Corsi, P. *More Is Better: Activators and Repressors from the Same Gene*. *Cell* **66**, (1992).

5. Koenig, R. J. *et al.* Inhibition of thyroid hormone action by a non-hormone binding c-erbA protein generated by alternative mRNA splicing. *Nature* **337**, 659–661 (1989).

6. Love, J. E., Hayden, E. J. & Rohn, T. T. Alternative Splicing in Alzheimer's Disease. *J. Park. Dis. Alzheimer's Dis.* **2**, (2015).

7. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–13 (2008).

8. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* **102**, 11–26 (2018).

9. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

10. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

11. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

12. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).

13. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Poznan, Poland)* **19**, A68-77 (2015).

14. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).

15. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**,

13 (2016).

16. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

17. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, (2016).

18. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5593-601 (2014).

19. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* **323**, 133–138 (2009).

20. Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics. Proteomics Bioinformatics* **13**, 4–16 (2015).

21. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* (2015). doi:10.1016/j.gpb.2015.08.002

22. Choudhury, O., Chakrabarty, A. & Emrich, S. J. HECIL: A Hybrid Error Correction Algorithm for Long Reads with Iterative Learning. *Sci. Rep.* **8**, 9936 (2018).

23. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* (2018). doi:10.1101/gr.222976.117

24. Abdel-Ghany, S. E. *et al.* A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* **7**, 11706 (2016).

25. Wyman, D. & Mortazavi, A. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics* **35**, 340–342 (2019).

26. Gordon, S. P. *et al.* Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PLoS One* **10**, e0132628 (2015).

27. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci.* (2014). doi:10.1073/pnas.1400447111

28. Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708 (2016).

29. Tseng, E. Cupcake ToFU.

30. Sahlin, K., Tomaszkiewicz, M., Makova, K. D. & Medvedev, P. Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nat. Commun.* **9**,

4601 (2018).

31. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).

32. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

33. Humme, S. *et al.* The EBV nuclear antigen 1 (EBNA1) enhances B cell immortalization several thousandfold. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 10989–94 (2003).

34. Arvey, A. *et al.* An Atlas of the Epstein-Barr Virus Transcriptome and Epigenome Reveals Host-Virus Regulatory Interactions. *Cell Host Microbe* **12**, 233–245 (2012).

35. Bodescot, M., Perricaudet,' And, M. & Farrell2, P. J. *A Promoter for the Highly Spliced EBNA Family of RNAs of Epstein-Barr Virus*. *JOURNAL OF VIROLOGY* (1987).

36. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).

37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

38. Pearce, S. F. *et al.* Maturation of selected human mitochondrial tRNAs requires deadenylation. *Elife* **6**, e27596 (2017).

39. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* (2018). doi:10.1101/410183

40. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol.* **5**, R74 (2004).

41. Cembrowski, M. S. *et al.* Dissociable Structural and Functional Hippocampal Outputs via Distinct Subiculum Cell Classes. *Cell* **173**, 1280-1292.e18 (2018).

42. Keil, J. M., Qalieh, A. & Kwan, K. Y. Brain Transcriptome Databases: A User's Guide. *J. Neurosci.* **38**, 2399–2412 (2018).

43. Holly, A. C. *et al.* Changes in splicing factor expression are associated with advancing age in man. *Mech. Ageing Dev.* **134**, 356–366 (2013).

44. Gordon, F. E. *et al.* Increased Expression of Angiogenic Genes in the Brains of Mouse Meg3-Null Embryos. *Endocrinology* **151**, 2443–2452 (2010).

45. Laeremans, A. *et al.* AMIGO2 mRNA expression in hippocampal CA2 and CA3a. *Brain Struct. Funct.* **218**, 123–130 (2013).

46. Liddelow, S. A. *et al.* Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).

47. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).

48. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

49. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

51. Anvar, S. Y. *et al.* Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, 46 (2018).

CHAPTER 4

A single-cell, isoform-level survey of the developing mouse forelimb using deep long-read sequencing

Chapter 4

A single-cell, isoform-level survey of the developing mouse forelimb using deep long-read sequencing

4.1 Abstract

During embryonic development, the diverse tissues of the mouse forelimb form through a highly regulated sequence of cell migration and differentiation. Although gene expression changes during limb development have been well-characterized, the role of alternative splicing and 5'/3' end variation during this process remains poorly understood. Here, we use single-cell PacBio long-read sequencing to deeply interrogate isoform-level expression in 81 selected cells from the myogenic and immune lineages in the embryonic mouse forelimb. At a depth of ~90,000 reads per cell, we find that long-read gene and isoform-level measurements show good agreement with short-read data from the same cells and are largely able to recapitulate the relationships between cell types. While we did not observe clear isoform switching events across cell types thus far, we did identify isoform-level markers that evaded the gene analysis, as well as novel isoforms that were specific to a particular cell state. In addition, our long-read data allowed us to examine alternative transcription start site (TSS) and transcription end site (TES) usage on the level of individual mRNA molecules. From this, we learned that some genes consistently express transcripts with multiple different TSSs or TESs in the same cell. These results demonstrate the utility of long-read sequencing for performing quantitative, isoform-level studies of single cell transcriptomes.

4.2 Introduction

Development is a carefully coordinated process in which cells must migrate and divide over a controlled timeline to form complex structures and tissues. Successive changes in gene expression are essential to reshaping the identity and functions of cells along the way[1,2]. With the availability of RNA-seq, a high-throughput method for measuring gene expression, it has been possible to thoroughly quantify transcriptional changes in a variety of developmental settings[3,4]. More recently, the extension of RNA-seq to single cells (scRNA-seq) has enabled unprecedented glimpses into how individual cells make fate decisions[5–8].

While many studies have focused broadly on the role of gene expression during development, it is believed that alternative splicing also plays an important role[9,10]. Alternative splicing allows individual multi-exonic genes to produce mRNA transcripts with different combinations of exons[11]. The resulting 'isoforms' may have different functional properties depending on which protein domains they code for, or alternately, due to the regulatory sequence domains they contain[12,13]. Isoform proportions have been observed to change across developmental stages even as the overall gene expression level remains stable, particularly in the brain[14]. Consequently, differential expression of isoforms, not just genes, can profoundly affect the properties of cells during development.

The developing embryonic mouse forelimb is an interesting setting in which to examine isoform expression because it contains a variety of migrating and differentiating cell types. Limb formation is highly conserved across vertebrates[15]. The mouse forelimb

forms over the course of embryonic day E9.5 through E17.5, beginning as a bud structure containing primarily undifferentiated mesoderm cells and later harboring bone, muscle, and cartilage tissues[16]. Skeletal muscle in the limb is believed to form by successive waves of cell migration and differentiation[17]. Initially, early muscle precursor cells expressing the *Pax3* transcription factor as well as *Myf5* migrate from the somite to the limb, forming masses that will later become distinct muscle groups[15,17,18]. Inside these masses, the muscle precursors differentiate into myoblast cells, which express myogenic lineage genes such as *Myod1* and continue to express *Myf5* [15,18,19]. Myoblasts can proliferate or terminally differentiate into myocytes, which are characterized by expression of genes such as *Myod1, Myog*, *Acta1*, and *Myh3*[20–22]. Early myocytes are mononuclear, but they later fuse into large, multinucleated myotubes, which organize into muscle fibers[20].

Recently, He and Williams et al. conducted an in-depth, single-cell study of the developing mouse forelimb using short-read single-cell RNA-seq[23]. Their analysis identified both known and new candidate cell types in the limb bud across developmental timepoints, and revealed regulatory circuits involved in various stages. A logical next step is to understand isoform-level expression in these cell types. Here, we use PacBio long-read technology to deeply sequence the full-length transcriptomes of 81 single cells isolated from five cell populations of interest in the developing mouse limb bud. We characterize and quantify the isoforms present in each cell using the TALON long-read pipeline from Chapter 3, identifying gene and isoform-level markers by cell type. We also use the TALON outputs to identify alternative 5' and 3' sites on a single-molecule basis. Overall, we demonstrate that it is feasible to quantify isoform-level expression in single cells using

long-read technology, and that these can provide insights that would be challenging to obtain from short reads.

4.3 Results

4.3.1 Single cell gene and isoform detection with PacBio long-read sequencing

Although single-cell sequencing has revolutionized the study of rare and heterogeneous cell types, characterizing single cells with a relatively low-throughput technology such as PacBio still poses a significant challenge. Notably, our populations of interest (muscle precursors, myoblasts, myocytes, EMPs, and macrophages) are extremely rare in the embryonic mouse limb bud overall. For instance, the largest of these lineages, the muscle precursors, are estimated to account for only 5.8% of limb bud cells[23]. Given the expense involved in long-read sequencing and the high read depth needed to characterize isoform-level expression, it is not feasible to sequence an indeterminate number of cells from the limb bud with PacBio and hope that enough cells of interest will be represented. Therefore, we devised a screening approach to allow targeted long-read study of selected cells **(Figure 4.1)**. This involved sequencing full-length, single-cell libraries with the comparatively cheaper short-read scRNA-seq first to assign cell type identities based on canonical gene markers, and then selecting libraries for long-read profiling based on these assignments. Using this approach, we performed PacBio long-read sequencing on 25 muscle precursors, 25 myoblasts, 20 myocytes, 5 EMPs, and 8 macrophages **(Figure 4.2,4.3a)**. Of these, two myocytes were later removed for quality control reasons.

Focusing on the relatively small total of 81 cells allowed us to achieve deep gene and isoform-level coverage, which is necessary to confidently detect quantitative isoform

differences between cells. A median of 92,150 reads per cell passed our quality controls and were used for gene and isoform calling in the TALON pipeline **(Figure 4.3b)**. After annotation, a median of 5,724 GENCODE genes were detected per cell, while a median of 5,123 known GENCODE isoforms were detected **(Figure 4.4a-b)**. Note that the gene count is higher than isoforms in part because novel isoforms can contribute to the detection of known genes. We observed a trend in which cell types considered to be less differentiated (i.e. muscle precursors, EMPs) tended to have a greater number of known genes and isoforms detected than more differentiated cell types (i.e. myocytes, macrophages). This effect was not satisfactorily explained by depth of sequencing alone **(Figure 4.5)**. Previous works have suggested that as cells differentiate, they devote increasing transcriptional resources to a more speciailized set of genes, leading to less diversity in the transcriptome [24]. This is one possible explanation for what we see in our data, but further investigation is needed.

We next compared our long-read gene detection to that of matched short-read scRNA-data for each individual cell. The number of genes detected was strongly correlated across the platforms (Spearman rho = 0.71), but as expected, short reads detected a greater number of genes per cell, likely due to higher sequencing depth **(Figure 4.6a)**. We also computed the gene expression correlation for each cell across the long and short-read platforms **(Figure 4.6b)**. The median Spearman rho across all cells was 0.65, indicating good agreement between the technologies. No particular trend was observed between the Spearman rho and the PacBio read count, indicating that all of our cells achieved enough long-read coverage to be comparable to the short-read measurements.

The TALON pipeline also allowed us to group novel isoforms into categories based on how they differed from known ones. After filtering to remove likely artifacts (see Materials and Methods), we detected a median of 1,482 novel isoforms per cell **(Figure 4.7a)**. The incomplete splice match (ISM) category was the most common novelty type, followed by the novel in catalog (NIC) category **(Figure 4.7b)**. Overall, the breakdown of novelty types was similar across the cells independent of type. Over 90% of the filtered reads came from known isoforms **(Figure 4.7c)**.

4.3.2 Clustering single cells based on gene expression

After obtaining gene and isoform-level counts from TALON on a per-cell basis, we used these as input to Scanpy[25] in order to cluster the cells and call gene and isoform markers. We performed the gene-level analysis first. We log-normalized the gene expression counts by cell and then performed dimensionality reduction via a principle component analysis (PCA). The Leiden algorithm was used to cluster the cells from the PCA, then we visualized the resulting clusters in two dimensions using the uniform manifold approximation and projection (UMAP) technique **(Figure 4.8a)**. For the most part, the cells clustered by known cell type identity, indicating that long-read derived gene expression measurements were sufficient to recapitulate the cell type composition of the data **(Figure 4.8b)**. Leiden clusters 0, 2, and 4 clearly corresponded to the muscle precursor, myocyte, and macrophage groups, respectively. Interestingly, the myoblast cells were divided across Leiden clusters 1 and 3, which we denoted as myoblast.1 and myoblast.2, respectively. Although the EMP cells were co-localized on the UMAP, the Leiden clustering failed to assign an EMP-specific cluster regardless of the parameter regime, conflating them instead

with the muscle precursors. This is likely due to the very small number of EMPs (5 cells) relative to the other cell types. We also superimposed the sequencing batch information of each cell onto the UMAP to visualize how the different batches segregated in two dimensions **(Figure 4.8c)**. The batches are well-distributed across the different clusters and do not seem to account for the substructure in the myoblast populations.

After computing cell clusters using Scanpy, we proceeded to call markers for each group. We identified 62 significant marker genes for the myocyte cluster, 65 for the macrophages, 22 for muscle precursor, 34 for myoblast.1, and 15 for myoblast.2 **(Table 4.1-4.4)**. Encouragingly, we found canonical cell type markers among the genes for each group, including *Pax3* for muscle precursors, *Msc* for myoblasts, *Myog* for myocytes, and *Spi1* for macrophages **(Figure 4.9)**. We also observed expected expression patterns for a set of well-known myogenic markers **(Figure 4.10).** For instance, *Pax3* was expressed mainly in muscle precursors, along with a handful of myoblasts. *Myf5* was pervasively expressed in muscle precursors and myoblasts, while *Msc* was largely specific to the latter group. *Des* and *Myod1* were expressed beginning in the myoblast stage but were more highly expressed in the myocytes. *Myog* was ubiquitously expressed by the myocytes and also appeared in a few myoblasts. A small number of myocytes dually expressed *Acta1* amd *Myh3*, which are considered later-stage myocyte markers.

Previous works have suggested that temporally distinct populations of myoblasts ultimately give rise to fast, mixed fast/slow, and slow-twitch muscle fibers[19,26,27]. Embryonic myoblasts emerge around day E10.5 and are believed to be precursors to

primary muscle fibers, which provide a scaffold for the developing muscle and tend to conform to the slow-twitch phenotype[28]. Fetal myoblasts appear later (~E14.5), differentiating into secondary myotubes that form fast-twitch fibers[28,29]. Because the single cells in our data were pooled from timepoints spanning E10.5 through E15.5, it is possible for the myoblast cells to belong to either group. We therefore sought to better understand the differences between the myoblast.1 and myoblast.2 clusters in our data, and to examine whether the groups corresponded at all to the embryonic and fetal subpopulations. Cells in the myoblast.1 cluster expressed more of the *Junb*, *Notch1*, and *Lmna* genes, which have been previously describedas enriched in fetal myoblasts **(Figure 4.11)**[19]. Meanwhile, myoblast.2 cells expressed more *Pax3*, *Met*, and *Myf5*, which have been observed at higher levels in embryonic myoblasts[19]. However, the results were less conclusive for other markers such as *Msc*, *Tcf15*, and *Prkcq*. Tentatively, it is possible that the myoblast.1 and myoblast.2 clusters correspond to the fetal and embryonic myoblast populations, but more evidence is needed. A possible confounder is the fact that the myoblast.1 cluster contains a handful of misclassified macrophage and EMP cells.

4.3.3 Examining splice isoform differences by cell type

Having established that long-read gene expression measurements could recapitulate cell type relationships in the single-cell data, we next analyzed the isoforms. First, we tested whether the cell types clustered as expected based on isoform as well as gene expression. To focus on the transcripts least likely to be artifacts, we elected to run Scanpy on isoform counts from known, NIC, and NNC transcripts only. As in the gene analysis, the isoform-based clustering successfully captured the underlying cell population structure

(Figure 4.12a). Again, the myocytes and macrophages formed clearly distinct groups, corresponding to Leiden clusters 2 and 4, respectively **(Figure 4.12b)**. In an interesting departure from the gene analysis, the myoblasts were grouped in a single cluster, number 0, while the muscle precursors were split in two (Leiden clusters 1 and 3). The sequencing batches were well-distributed across the UMAP **(Figure 4.12c)**. We next identified 64 marker isoforms for the myocyte cluster, 69 for the macrophage cluster, 41 for the myoblasts, 42 for the muscle-pre.1 cluster, and 25 isoforms for the muscle-pre.2 cluster **(Figure 4.13; Table 4.5-4.8)**. 34 of the markers were novel NIC or NNC isoforms, which likely would have been difficult to call with short reads. Of the 241 total isoform markers, 114 belonged to genes that were called as markers in the previous gene analysis, and 127 were unique to the isoform analysis **(Table 4.9)**. Although we observed several cases of marker isoforms belonging to the same gene, all of these instances occurred within the same cluster rather than across cell types. For instance, four different *Tnnt1* isoforms were called as myocyte markers, and three *Msc* isoforms were called as myoblast markers.

Athough we did not find clear examples of isoform switching in our marker analysis, we nevertheless identified interesting examples of context-specific alternative splicing in the data at large. For instance, Rbm24 is a splicing factor known to regulate skeletal muscle by promoting inclusion of muscle-specific exons in its targets[30]. We detected known isoforms *Rbm24-201*, *Rbm24-203* and *Rbm24-204* in all three muscle cell types, and also identified a novel isoform (*Rbm24-ENCODEMT000167653*) that contained a new combination of known exons **(Figure 4.14)**. This novel transcript was exclusively expressed in myocyte cells and would have been challenging to identify from short reads.

Going forward, it would be interesting to further investigate the functional properties of this novel isoform relative to the others, and to ask whether the exon inclusion event has any effect on Rbm24's activity as a splicing factor.

4.3.4 Alternative 5' and 3' ends

Because PacBio is a single-molecule sequencing technology, it is possible to identify the 5' start and 3' ends for each transcript in addition to its splicing pattern, a task that is very difficult to perform with short-read data. This led us to ask whether we could observe any interesting trends in the single cells with respect to transcript differences at the 5' and 3' ends. We first called transcription start sites (TSSs) directly from the long reads that passed our isoform filters **(Figure 4.15a)**. This was done on a per-gene basis by selecting the start positions of each long-read alignment that was annotated to the gene and collapsing these sites into ranges (see Materials and Methods for details). Since long reads are subject to artifacts that can generate misleading starts and ends, we also sought to assign evidence levels to the TSSs. For instance, we compared each TSS to known start sites from the GENCODE annotation as well as FANTOM CAGE peaks. We found that of 39,209 total TSSs called from the filtered transcript models, 15,600 TSSs were supported by > 1 read and had GENCODE support, CAGE support, or both. This total can be subdivided to 11,638 known GENCODE TSSs and 3,962 CAGE-supported novel TSSs. A median of 4,303 GENCODE/CAGE-supported TSSs were detected per cell **(Figure 4.15b)**. We next plotted the number of known splice isoforms and supported TSSs detected for each gene in each individual cell **(Figure 4.15c)**. Although most genes had at most one supported TSS and one known isoform detected at a time in a single cell, there were many exceptions. For

instance, four distinct TSSs of the *Hes1* transcription factor were found in a single myoblast and myocyte cell, and a further 6 muscle-precursors, 4 myoblasts, and 1 myocyte cell expressed more than one *Hes1* TSSs at the same time **(Figure 4.15d)**. Interestingly, many of these distinct TSSs came from transcripts that otherwise shared the same set of exons. The *Hnrnpf* RNA binding protein gene also displayed a striking amount of TSS heterogeneity on the single-cell level. When only known isoforms were considered, all but one cell in the muscle precursor, myoblast, and myocyte categories expressed two or more *Hnrnpf* TSSs, the max being 4 per cell **(Figure 4.15e)**. When we included reads that were annotated to novel NIC and NNC isoforms as well, a maximum of 5 TSSs were detected in 21 individual cells **(Figure 4.15f)**. In both versions, the muscle precursor population had a significantly higher degree of TSS diversity in *Hnrnpf* than was observed in the myoblast and myocyte cells. The TSSs differences discussed here would have been difficult to discern from short-read RNA-seq data alone and represent a case where single-molecule long reads can make a substantial contribution.

The next step was to identify and compare transcription end sites (TESs) in the single cells. Candidate TESs were called in a similar manner to the TSSs, though the evaluation was somewhat different **(Figure 4.16a)**. For each TES, we compared the candidate to known GENCODE end sites and looked for the presence of a canonical poly-(A) motif towards the end of the sequence. Of the 30,671 total TESs called, 20,347 had GENCODE support, poly(A) motif-support, or both. We detected 12,200 known GENCODE TESs and 8,147 poly(A)-supported novel TESs. A median of 4,370 GENCODE/poly(A) motif-supported TESs were detected per cell **(Figure 4.16b)**. As in the case of the TSSs, most cells

detected one TES and one splice isoform at most per gene **(Figure 4.16c)**. However, exceptions included the tropomyosin 1 (*Tpm1)* gene – cells from the three muscle populations were consistently found to simultaneously express two or more validated TESs in mucle precursor, myoblast, and myocyte cells **(Figure 4.16d).** Muscle precursors contained significantly more *Prrx1* TESs per cell than the myocytes **(Figure 4.16e)**. Furthermore, the muscle precursor and myoblast cells expressed significantly more TESs of splicing factor *Srsf3* than did the myocytes **(Figure 4.16f).** This is intriguing because the expression level of this gene was relatively steady across all three cell types **(Figure 4.17a-b)**. This further illustrates the transcriptomic diversity that can be found on the level of single cells using long-read sequencing.

4.4 Discussion

Here, we used PacBio long-read sequencing to deeply profile the isoform-level transcriptomes of 81 individual muscle precursor, myoblast, myocyte, EMP, and macrophage cells from the developing mouse limb. Overall, our results demonstrate that it is possible to quantitatively measure gene and isoform expression as well as alternative TSS/TES usage directly from long reads on the single-cell level. We found that long read-based expression levels correlated well with short-read data from the same individual cells. In addition, long-read clustering analyses on both the gene and isoform level reconstituted known celltype groupings with respect to the three myogenic populations. In particular, the isoform expression analysis revealed over 200 isoform-level cluster markers, many of which contained novel splicing events. More than half of these isoforms came from genes that were not identified as gene-level cluster markers, suggesting that some differences

between the cell types may be driven on the level of particular isoforms rather than genes. An important next step will be to characterize the functions of differentially expressed isoforms in order to better understand how they may be contributing to the underlying biology of the cells they are expressed in.

While we did observe isoforms that were specific to particular cell populations, we were somewhat surprised that we did not observe any obvious isoform switching events in which two isoforms showed inverted expression patterns across celltypes. One possible explanation could be that not much isoform switching goes on in the particular cell types and stages that we chose to observe. Isoform switching is known to be an important developmental mechanism in tissues such as the brain, but the ground truth is less well understood in myogenesis. However, it is also possible that we were unable to detect these events as a result of our current sequencing depth and quantity of cells. One option would be to sequence additional cells from the original screen to approximately double the total. An important challenge of working with long read data is how to use reads that are annotated to potentially artifactual novel transcripts. In particular, the incomplete splice match (ISM) category presents a conundrum because these transcripts may represent 'lost' expression from a longer isoform, and it is often unclear exactly which one. Here, we chose to be conservative by including only known, NIC, and NNC transcripts in our isoform-level analyses, but it is possible that we missed some isoform expression patterns in doing so. It is also worth noting that the Scanpy package was developed and optimized for short-read gene expression analysis in large single cell datasets, rather than for long-read, isoform-based counts. Since signal dropout for isoforms is likely to be even more severe than on the

gene level, computational methods may need to be tailored more specifically to help find differential isoform trends hidden in the noise.

There is one final limitation to consider. We did not add unique molecular identifiers during PacBio library preparation, meaning that it was not possible to detect and remove PCR duplicate reads. When working with single cells, there is always a concern that the large amounts of amplification needed to build a library can give rise to runaway amplification artifacts that distort expression measurements downstream. While we cannot completely rule out this scenario in our data, we do not believe that it invalidates the general conclusions we have drawn.

In the third-generation sequencing era, methods that combine single-molecule sequencing with the resolution of single cells will offer an unprecedented look at the transcriptional diversity underpinning different cell states. The methods we have described here can be extended and applied in a variety of biological settings. We anticipate that long-read, single-cell transcriptome profiling will be highly useful in advancing our understanding of disease, particularly for neurological disorders where alternative splicing is known to play a role.

4.5 Materials and Methods

4.5.1 Library selection and long-read sequencing

Single cells dissected from the developing mouse limb bud (embryonic days E10.5, E11, E11.5 E12.5, E13, E13.5, E14, E14.5 and E15.5) were isolated using the Fluidigm C1

system and were built into Smart-seq full-length cDNA libraries with ERCC spike-ins as described in He, Williams, et al. 2020[23]. These libraries were first sequenced as 50 bp single reads on the Illumina Hi-Seq 2500 to a depth of about 1M aligned reads (cell barcodes were added at the tagmentation step). Cell type identities were called as described in He, Williams, et al. 2020[23]. Using these identities, a total of 83 cells were determined to belong to cell types of interest (muscle precursors, myoblasts, myocytes, erythroid myeloid progenitors, and macrophages) and were selected for further study by long-read sequencing. To prepare these single-cell cDNA libraries for PacBio sequencing, we added cell-specific barcodes via limited PCR. After pooling the bar-coded libraries in equi-molar proportions, a portion of the pool was reserved for size selection. Concentration of the pooled libraries was determined via Qubit, and the pool was then reconcentrated to > 45 ng/uL using Ampure XP SPRI beads. The pool was then bound to Sera-Mag beads for size selection, collected on a magnet and rinsed twice with 70% Etoh, and then eluted from the beads in EB. Sequencing was performed on a combination of the PacBio Sequel and Sequel II machines **(Table 4.10)**.

4.5.2 Running TALON pipeline on single-cell long-read data

The full workflow is shown in Figure 2. The CCS program was run on each raw library subreads BAM file to generate a high-fidelity consensus for each read (version 4.0.0, parameters: --skip-polish --min-length=10 --min-passes=3 --min-rq=0.9 --min-snr=2.5). The CCS reads were demultiplexed by cell barcode and had their adaptors removed by the Lima program (version 1.10.0, parameters: --same --split-bam-named --score-full-pass --ccs --dump-removed). At this point, each library was subdivided into separate fasta files for each

cell. Next, we ran a custom script (flip_reads.py) on each fasta file to identify and reverse-complement reads with a poly(T) sequence of at least 20 bp in the first 50 nucleotides of the read. The purpose of this step was to place reads in the correct strand orientation prior to aligning them to the mm10 + ERCC reference genome with Minimap2[31] (version 2.17, parameters -ax splice:hq -uf). After alignment, reads were corrected using TranscriptClean[32] (version 2.0.2, parameters: --canonOnly + defaults). Next, internal priming scores were computed for each read using the talon_label_reads utility (TALON[33] version 5.0, parameters --ar 20). At this point, we created a unified SAM file for each cell ID by concatenating all reads for each cell from different sequencing runs and libraries. We imposed a minimum read cutoff of 5,000, which led us to remove two cells with identifiers 20036_D11 and 19915_B9 (both myocytes).

A mouse TALON database was initialized from the GENCODE vM21 + ERCC annotations using the **talon_initialize_database** module from the TALON package (parameters: --l 0 --5p 500 --3p 300). To annotate the reads, we created a configuration file with a line for each cell and ran the **talon** module on this file along with the TALON database (parameters: --cov 0.9 --identity 0.8).

To perform long read quantification, transcript abundance matrices were extracted from the TALON databases using the **talon_abundance** module. We used the unfiltered abundance files for all gene-level expression analyses (omitting genomic transcripts). To filter isoforms, we required the isoform be a) known, or b) to appear at least once in a minimum of 5 cells of the same cell type. Reads with > 0.5 fraction As (as specified by **talon_label_reads**) were omitted when computing this read support. The union of the

resulting whitelists was used to generate filtered abundance files for transcript quantification (using **talon_abundance)**, as well as custom filtered GTF annotations (using **talon_create_GTF***)*.

4.6 Short-read gene quantification

Matching short-read RNA-seq data was available in the fastq format for each single cell. To quantify the short-read gene expression, we ran Kallisto[34] (version 0.43.1; parameters -b 100 –single -l 180 -s 20) on the fastqs for each individual cell, using a Kallisto index generated from the GENCODE vM21 transcriptome annotation plus ERCC spike-in sequences. Genes expressed at a level of < 1 TPM were not considered in further analysis.

4.7 Scanpy long-read gene analysis

Scanpy version 1.4.4.post1 was used. A single-cell gene expression matrix was constructed by summing all transcript counts per gene for each cell from the unfiltered TALON abundance file, excluding only the genomic transcript category. This matrix was used to initialize a Scanpy AnnData object. Gene expression values were normalized to 10,000 total counts per cell and log-transformed as recommended in the Scanpy documentation. Next, we called highly variable genes using Scanpy's scanpy.pp.highly_variable_genes function (parameters: n_top_genes = 3000, flavor = 'seurat', min_mean=0.0125, max_mean=3, min_disp=0.5). We performed PCA dimensionality reduction (scanpy.tl.pca) on the highly variable gene values scaled using scanpy.pp.scale (parameters: max_value=10). To select the number of PCs to use for subsequent analyses, we computed the number of PCs needed to explain 90% of the

variance. The result was 36 PCs. We computed a neighborhood graph of the PCA-processed cells using scanpy.pp.neighbors (n_neighbors=9, n_pcs=36). Then, we performed Leiden clustering using scanpy..tl.leiden. We used scanpy.tl.umap and scanpy.pl.umap to embed the graph in two dimensions. Cluster markers were called using sc.tl.rank_genes_groups (method='wilcoxon', use_raw = True, corr_method= 'benjamini-hochberg'), then filtered using sc.tl.filter_rank_genes_groups (min_in_group_fraction= 0.25, min_fold_change= 1, max_out_group_fraction= 0.5).

4.8 Scanpy long-read isoform analysis

This analysis was conducted largely the same way as the Scanpy gene analysis, though on a different input. Rather than computing gene-level expression, we initialized the Scanpy AnnData object on an expression matrix derived from the filtered TALON abundance file, and elected to keep only transcripts from the known, NIC, and NNC categories. This was done because these are the novelty types with the most supporting evidence. The number of PCs needed to explain 90% of the transcript variance was 36. After marker calling, the markers were filtered using sc.tl.filter_rank_genes_groups (min_in_group_fraction= 0.25, min_fold_change= 1, max_out_group_fraction= 0.5).

4.9 TSS and TES isoform analysis

To call long-read transcription start sites (TSSs), we took the 5’ end start site of each PacBio read annotated to a transcript model that passed the TALON filter and internal priming cutoff. These were recorded in the TALON read annotation output file. Next, for each gene individually, we merged the read starts such that consecutive members of the

same TSS group were at most 50 bp apart. To filter the TTSs, we required them to be located within 10 bp of either an annotated GENCODE TSS or a FANTOM5 CAGE peak, and also to be supported by at least two reads. The CAGE peaks were downloaded in the BED format from the following: https://fantom.gsc.riken.jp/5/datafiles/reprocessed/mm10_latest/extra/CAGE_peaks/mm10_fair+new_CAGE_peaks_phase1and2.bed.gz.

We used a similar approach to call long-read transcription start sites (TESs). We took the 3' end start site of each PacBio read annotated to a transcript model that passed the TALON filter and internal priming cutoff, merging consecutive sites for each gene that were within 100 bp of each other. To filter the TESs, we looked for an annotated GENCODE end within 10 bp of the TES and/or the presence of a canonical poly(A) motif in the last 35 bp of the TES interval as described in Anvar *et al*. 2018[35]. TESs that were supported by fewer than two reads were classified as unsupported.

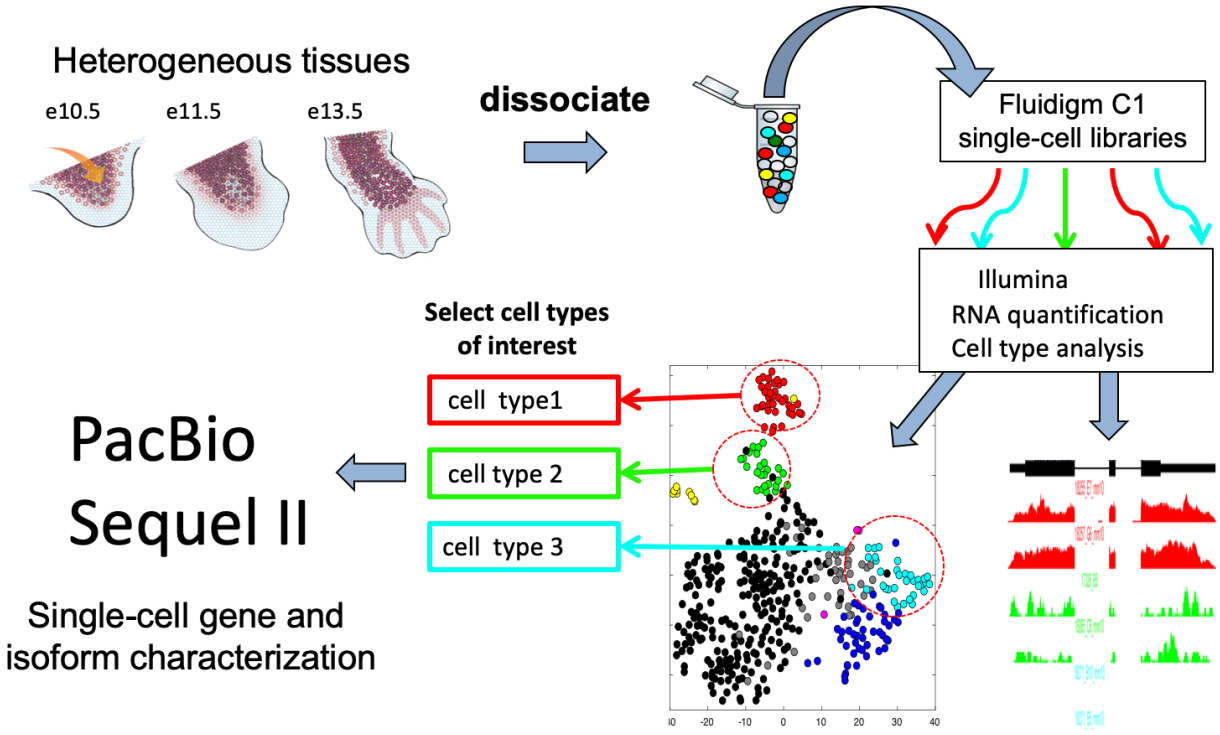

Figure 4.1. Experimental design for selecting cells of interest for long-read profiling. Single-cell SMART-seq libraries are prepared from cells isolated from the mouse limb bud and sequenced using short reads. These data are used to determine the identity of each cells. Cells belonging to lineages of interest are then sequenced on the PacBio long-read platform.

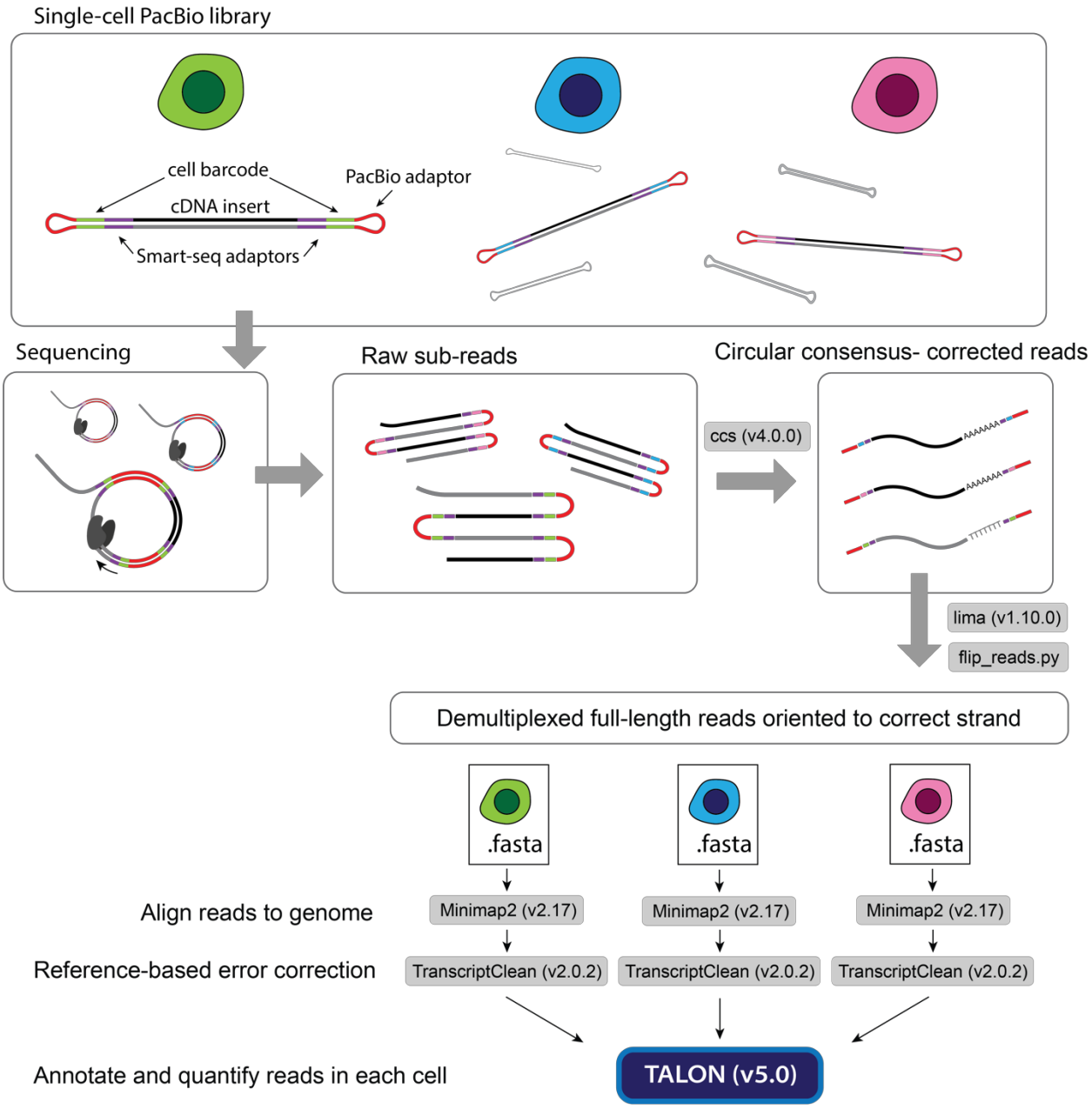

Figure 4.2. Single-cell barcoding scheme and computational workflow. Cell-specific barcodes and PacBio circular adaptors were added after the initial Smart-seq library prep. After sequencing and standard circular consensus correction, reads were demultiplexed and partitioned by cell barcode. The reads for each cell were then processed separately up until the TALON annotation step.

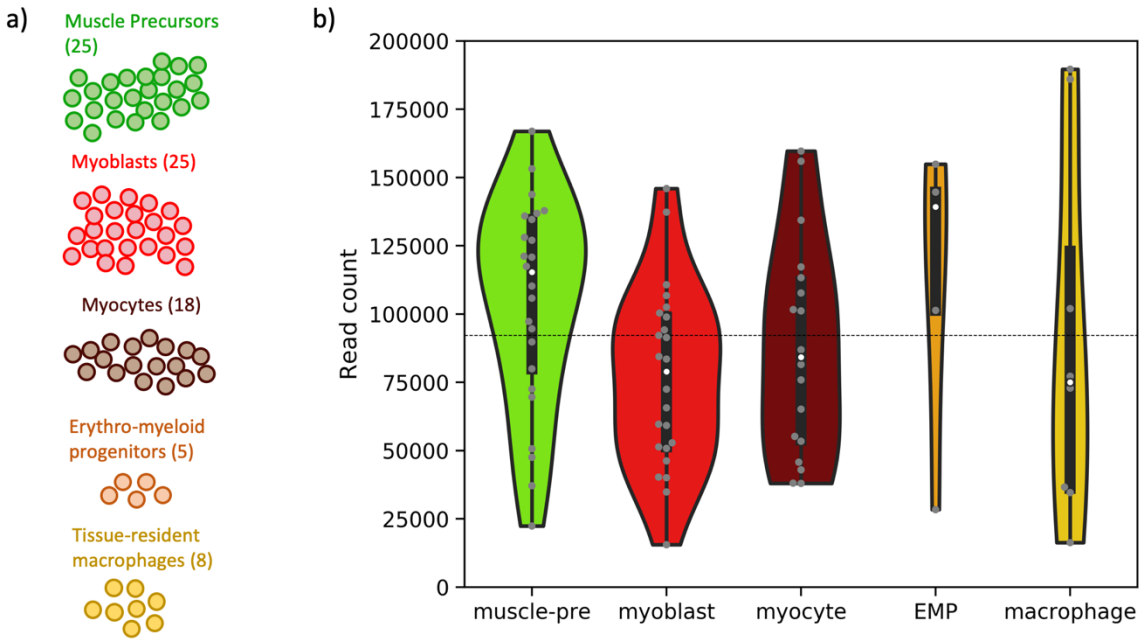

Figure 4.3. Deep sequencing of limb bud cells with PacBio. a) Number of cells sequenced per cell type. **b)** Number of reads obtained per cell at the annotation (TALON) stage of the analysis. The median value across all cell types was 92,150 reads.

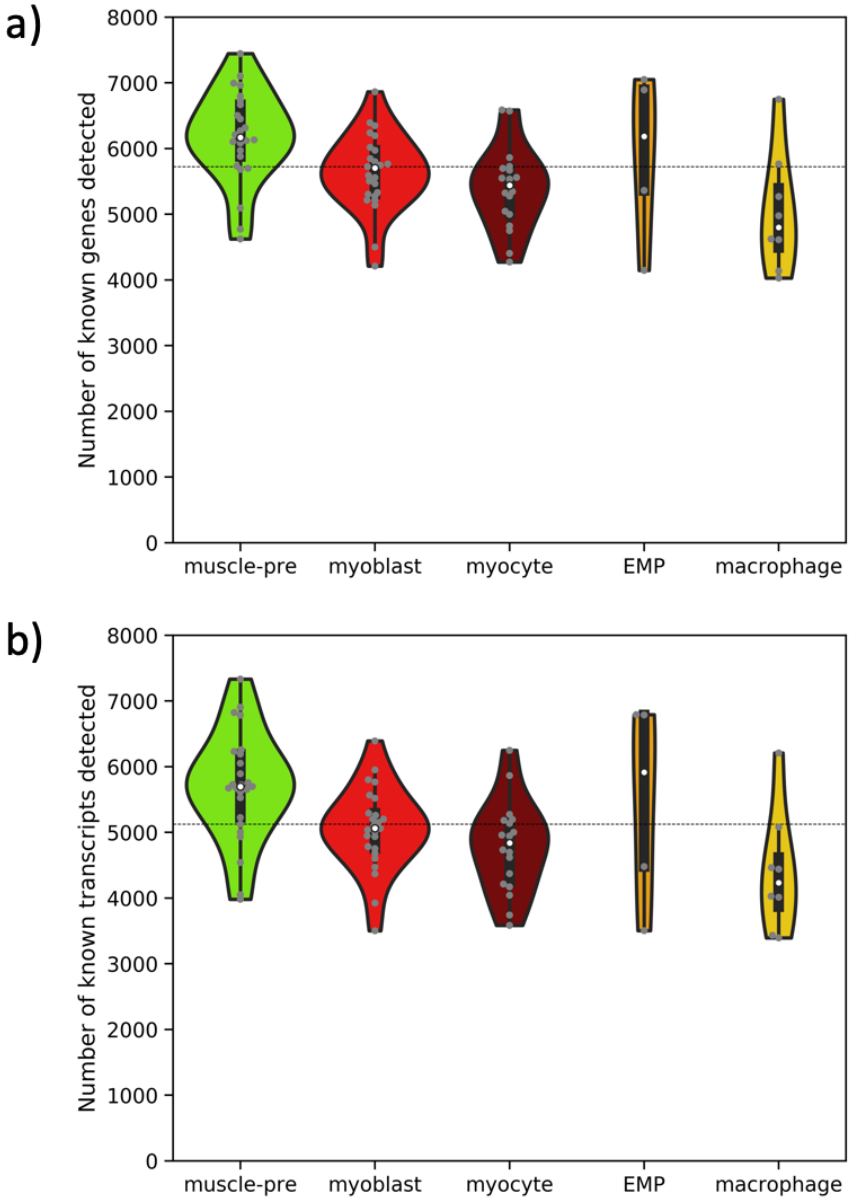

Figure 4.4. Long-read gene and isoform detection in single cells. a) Number of known GENCODE genes detected per single cell across cell types. Median value of 5,724 genes per cell. **b)** Number of known GENCODE isoforms detected per single cell across cell types. Median value of 5,123 isoforms per cell.

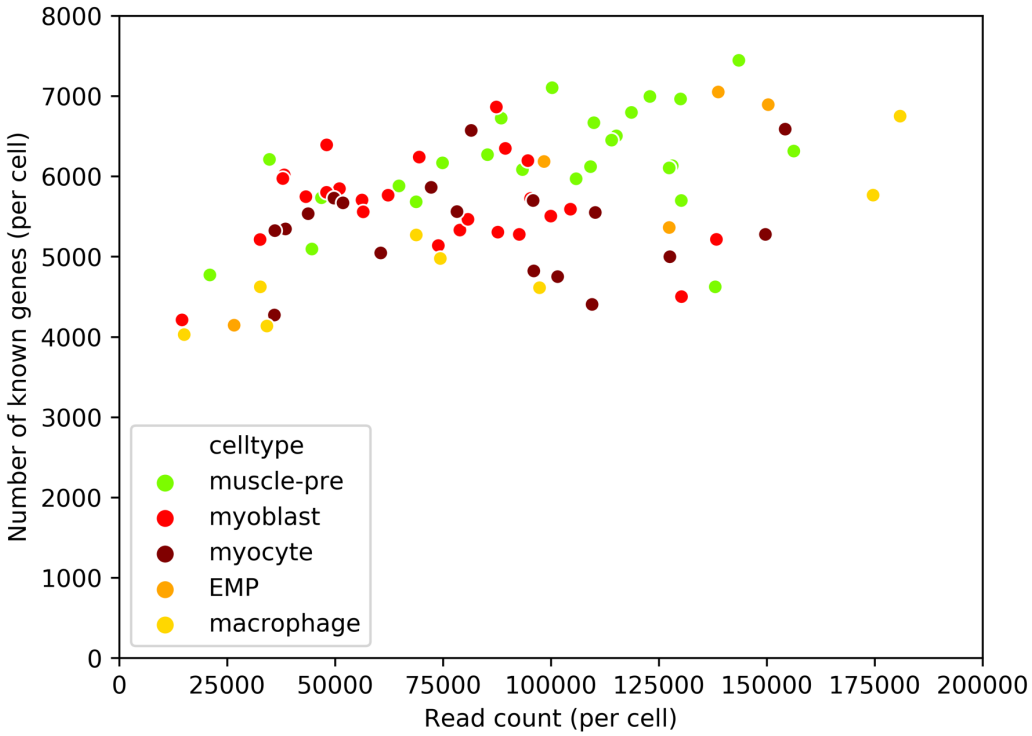

Figure 4.5. Number of genes detected per cell versus of long-read count. The identity of each cell is indicated by its color.

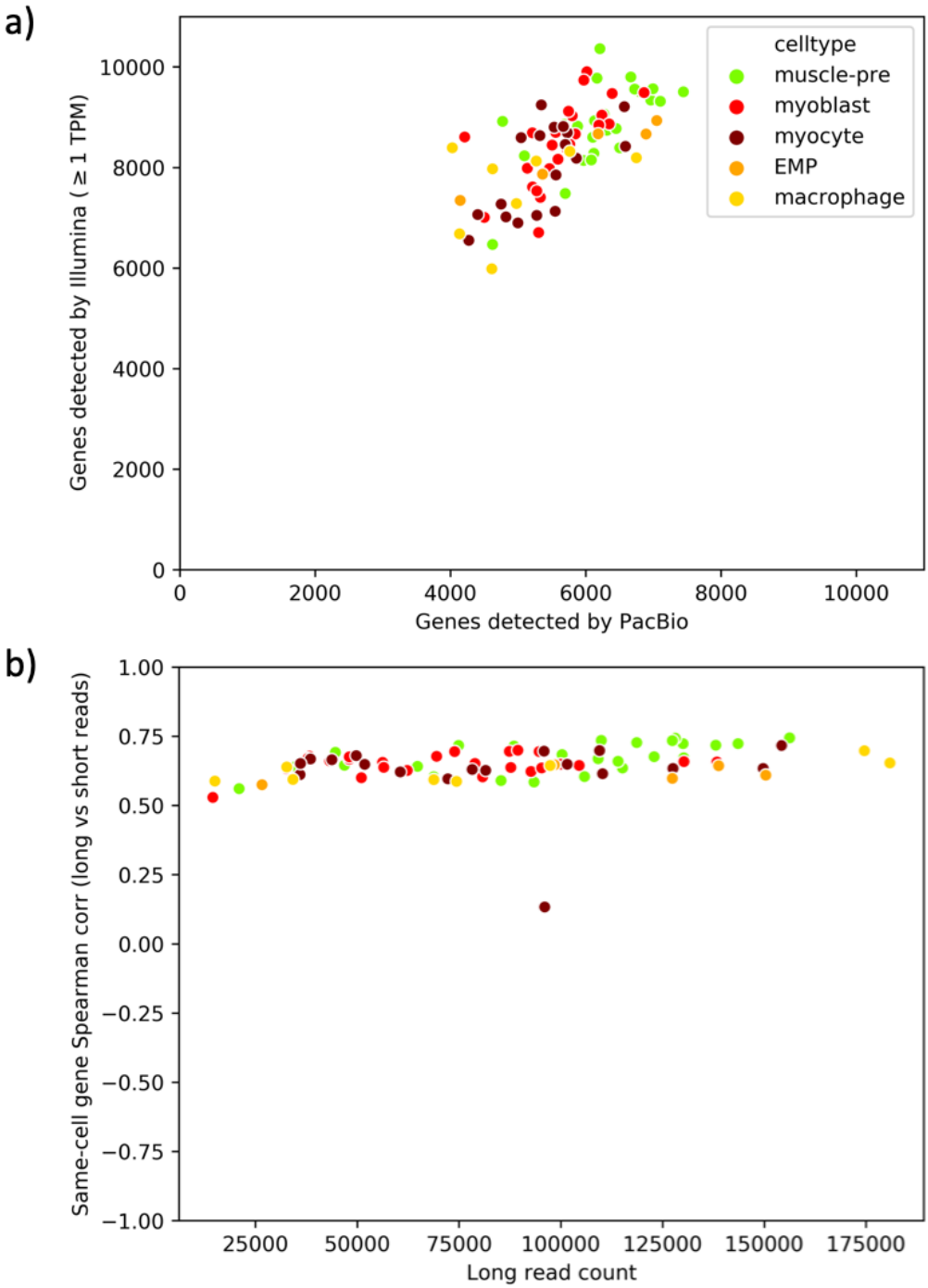

Figure 4.6. Comparison of long and short-read gene expression measurements in single cells. a) Number of genes detected in the same cell by PacBio long-read and Illumina short-read sequencing. Libraries were built using the Fluidigm C1 platform. **b)** Gene expression Spearman rho computed for each cell across long and short-read platforms, plotted against the number of long reads sequenced per cell.

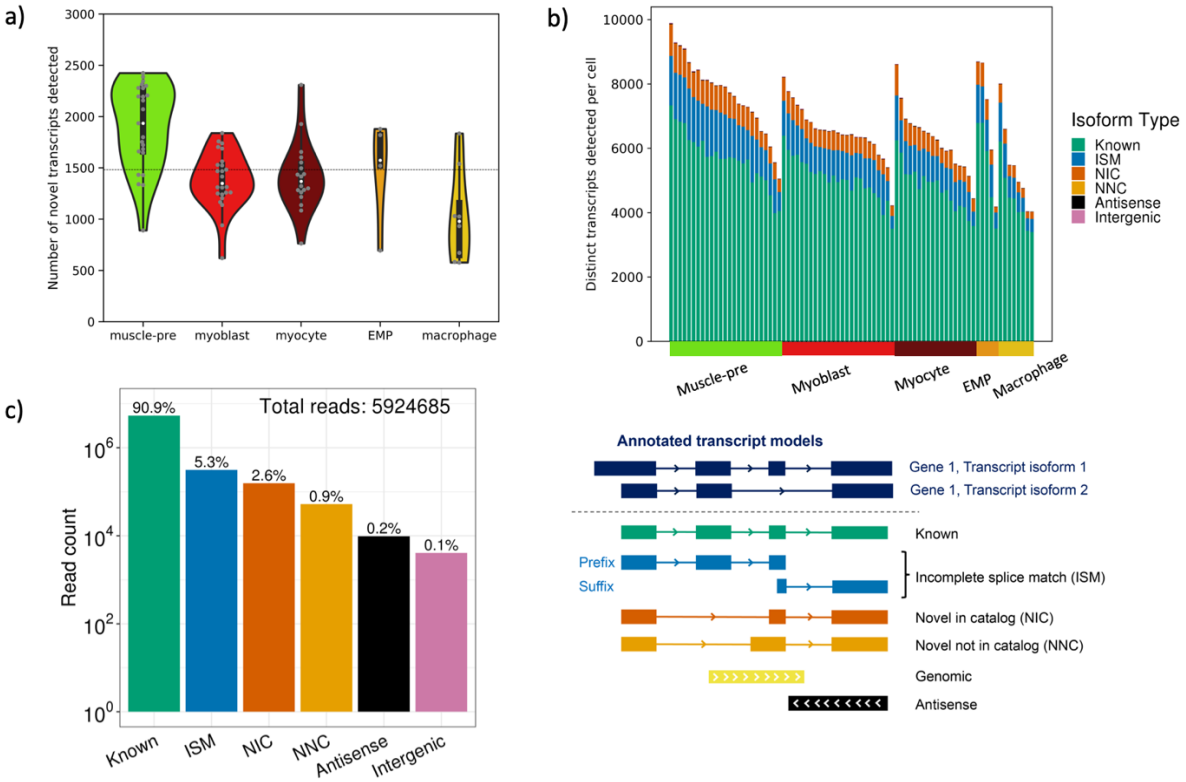

Figure 4.7. Novel transcripts identified by TALON from PacBio single cells. a) Number of novel transcript models detected per cell by TALON (after filtering). **b)** Novelty breakdown of isoforms detected in each single cell (after filtering). **c)** Novelty category assignments of filtered reads (pooled from all cells). A diagram of the novelty categories is shown alongside.

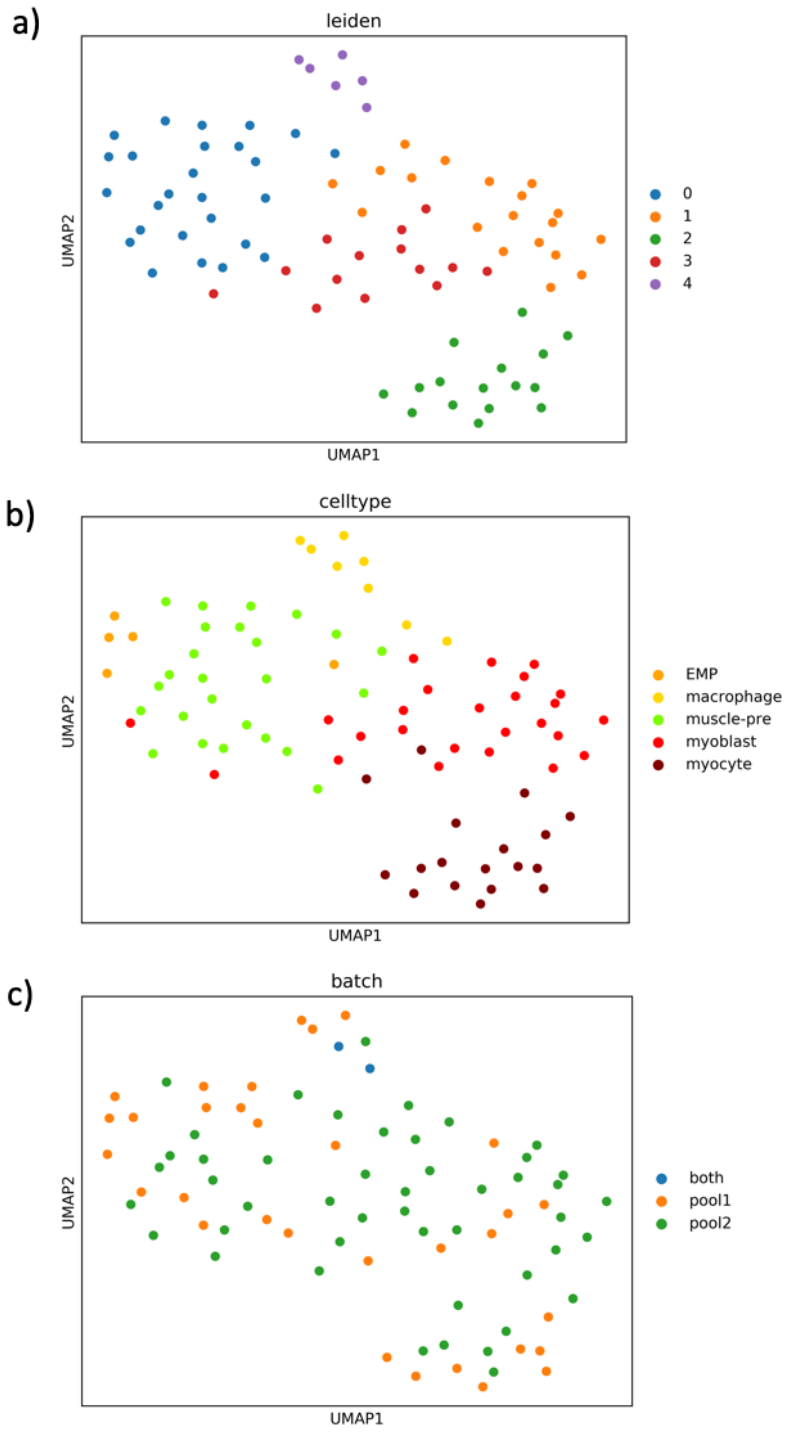

Figure 4.8. UMAP visualization of single cells based on long-read gene expression. a) Clusters derived from Leiden neighborhood graph. **b)** Cells labeled by the cell identity that was assigned by the short-read screening process. **c)** Cells labeled by Sequel II sequencing batch. 'Pool 1' refers to cells that were sequenced at HudsonAlpha, 'Pool 2' refers to cells sequenced at UC Irvine, and 'both' describes cells that were sequenced at both facilities.

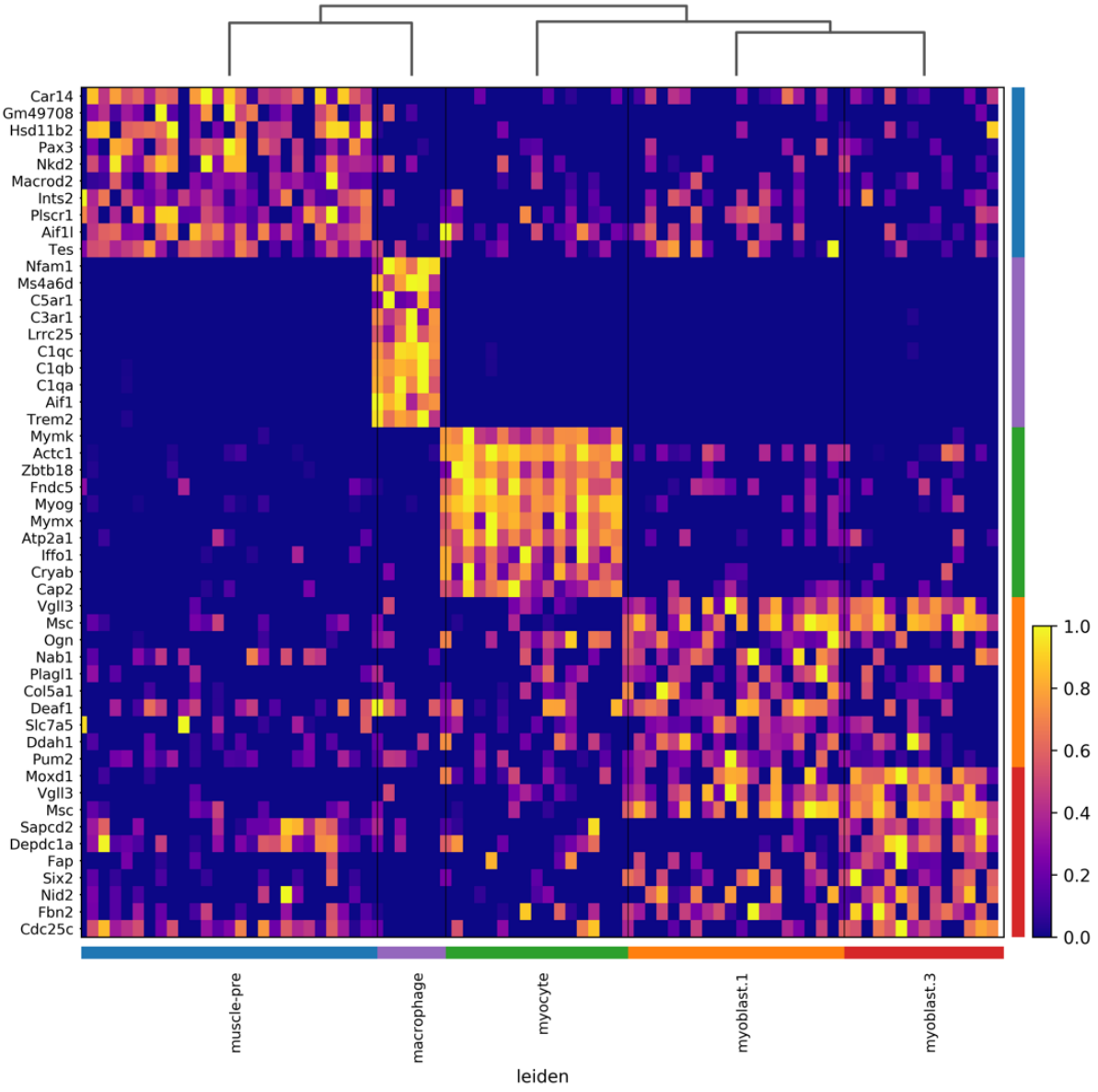

Figure 4.9. Expression of gene markers identified for Leiden clusters. The top 10 gene markers (as ranked by Scanpy) are shown for each group.

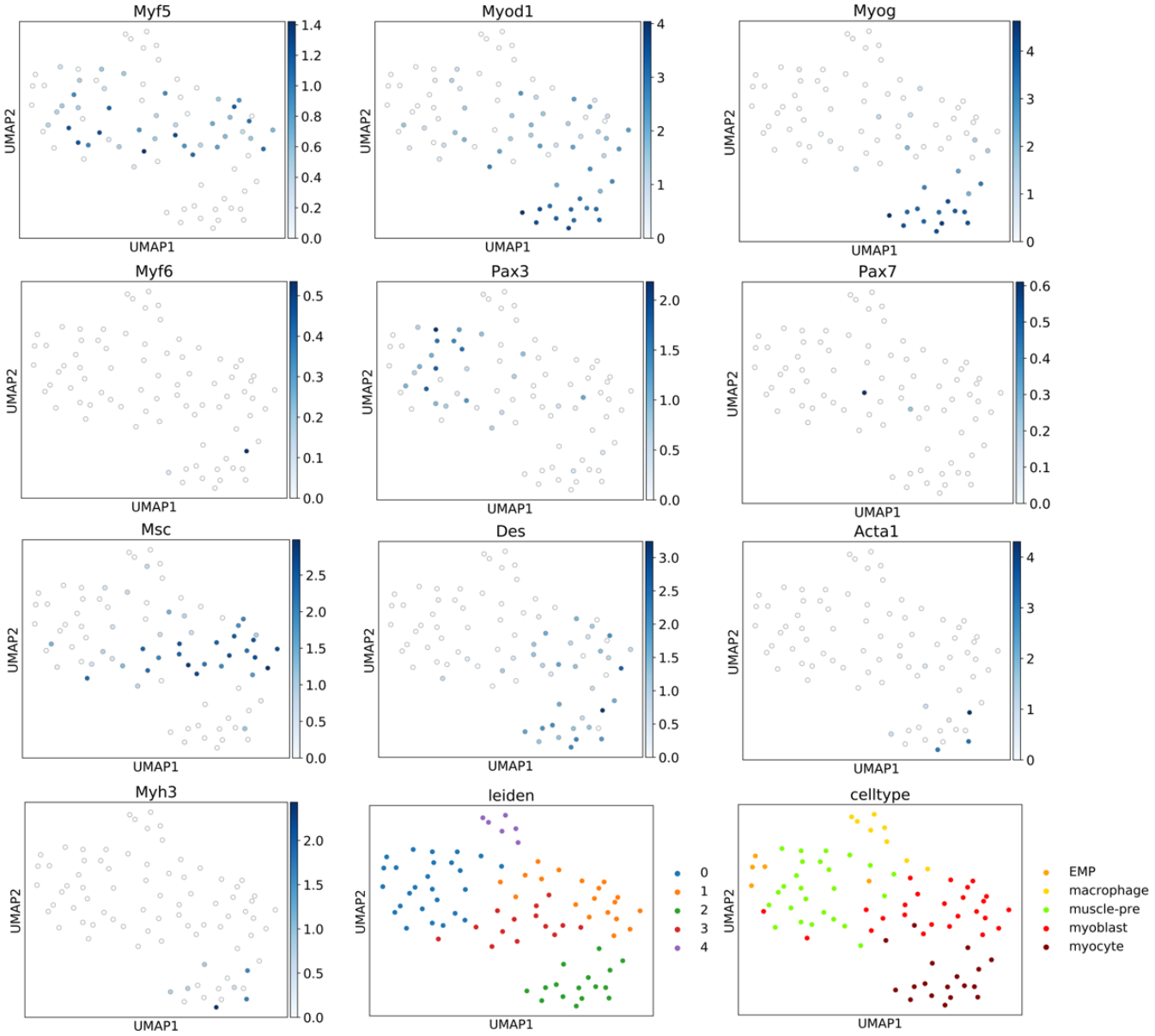

Figure 4.10. Expression of selected myogenic lineage marker genes in each cell.
Expression levels are overlaid on the Scanpy gene UMAP.

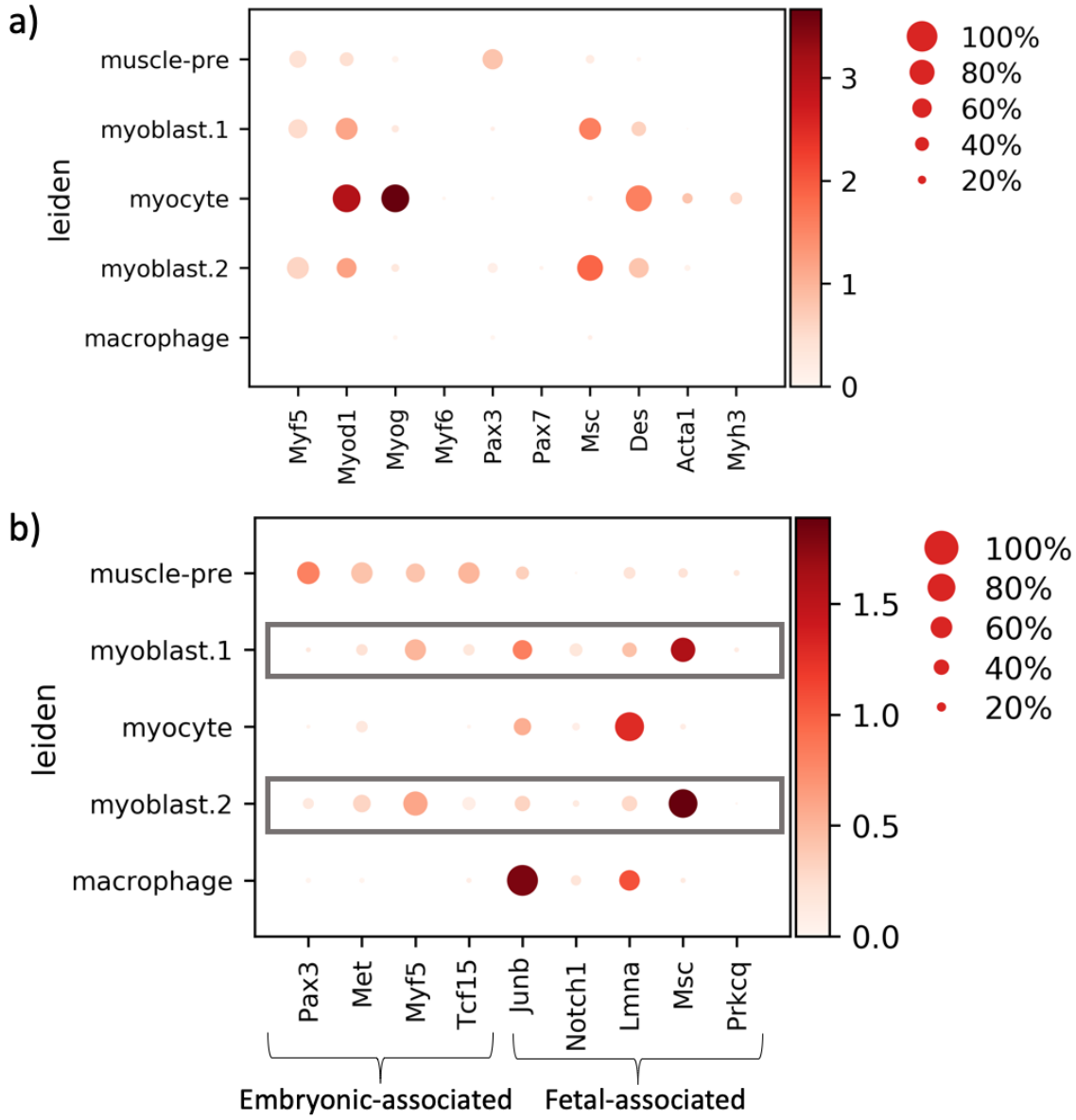

Figure 4.11. Mean PacBio expression level and fraction of cells expressing selected genes in each Leiden cluster. a) Myogenic lineage genes. **b)** Genes previously described as expressed in embryonic and fetal myoblast cells.

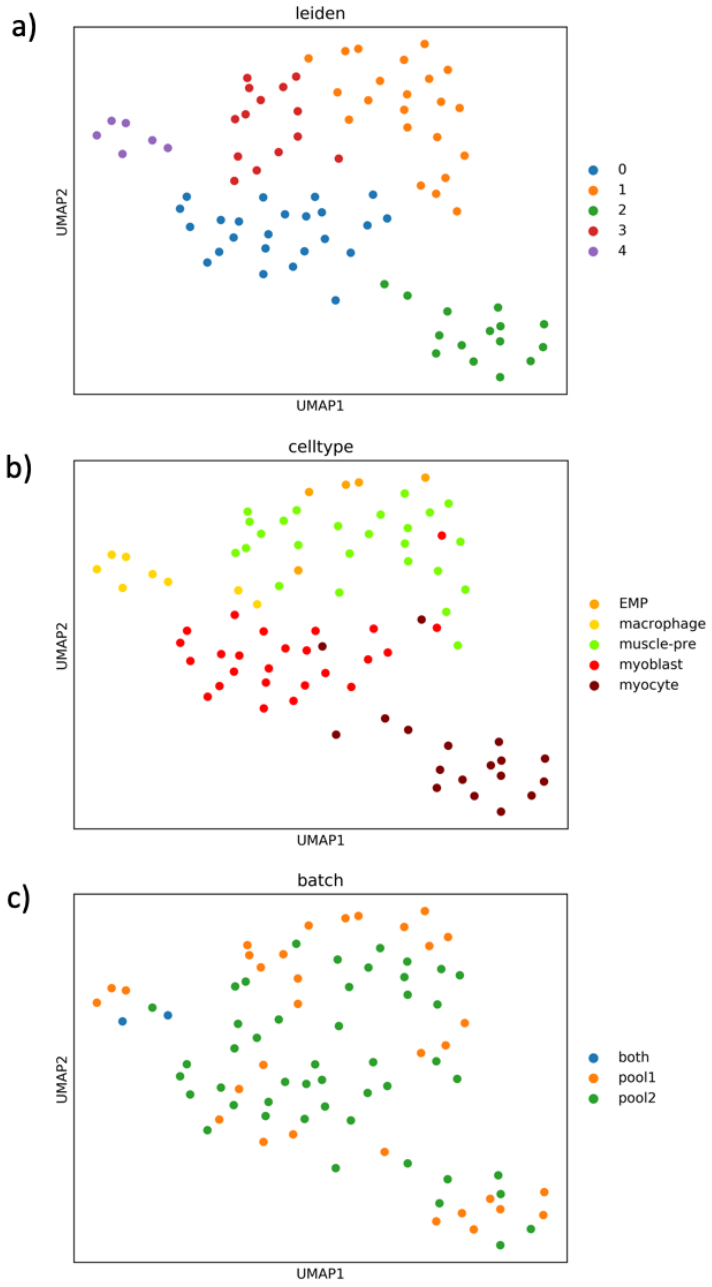

Figure 4.12. UMAP visualization of single cells based on long-read isoform expression (Known, NIC, and NNC only). a) Clusters derived from Leiden neighborhood graph. **b)** Cells labeled by the cell identity that was assigned by the short-read screening process. **c)** Cells labeled by Sequel II sequencing batch. 'Pool 1' refers to cells that were sequenced by the HudsonAlpha facility, 'Pool 2' refers to cells sequenced at the UC Irvine facility, and 'both' describes cells that were sequenced at both facilities.

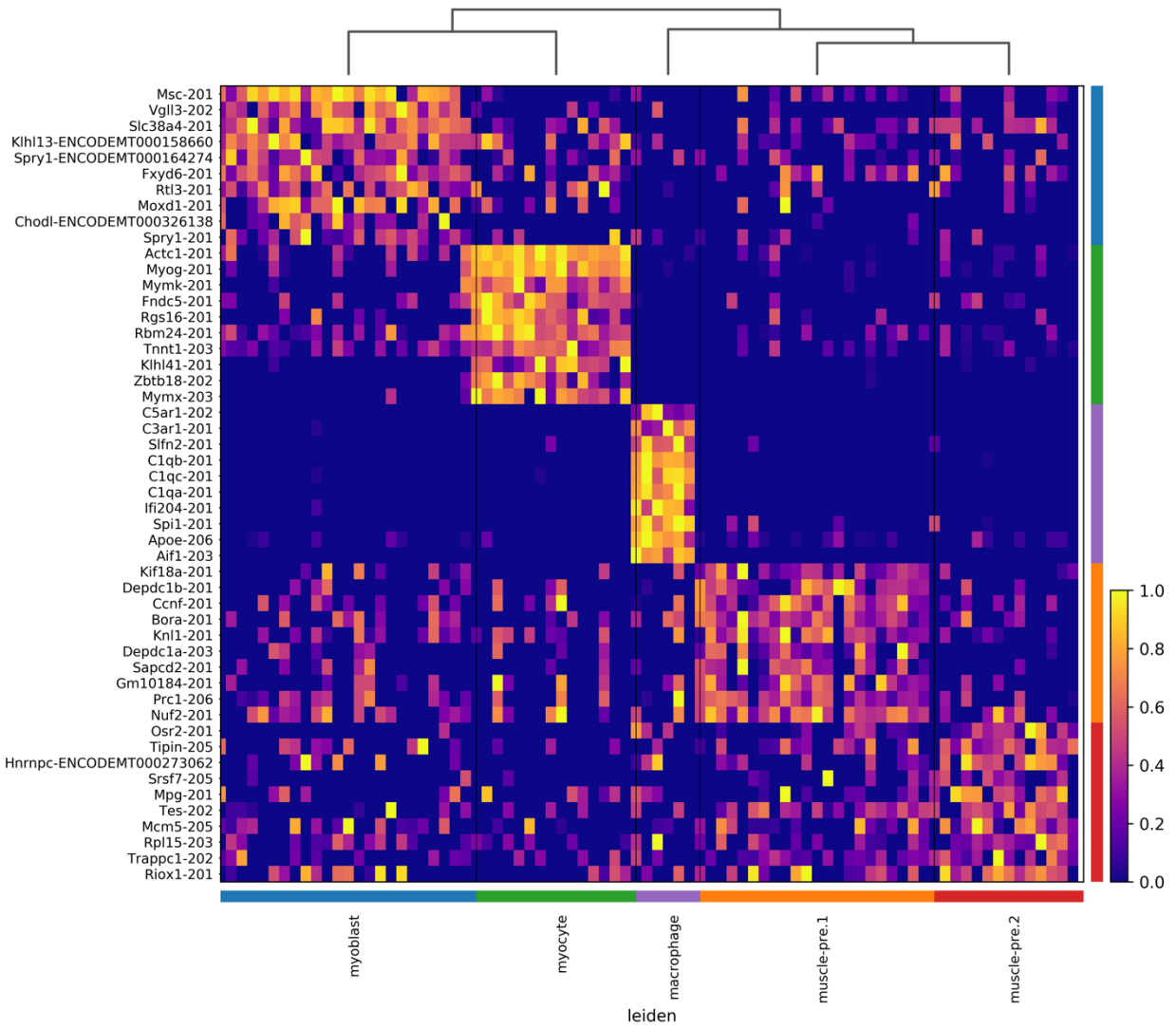

Figure 4.13. Expression of isoform markers identified for Leiden clusters. The top 10 gene markers (as ranked by Scanpy) are shown for each group. Isoforms with 'ENCODE' in the identifier are novel.

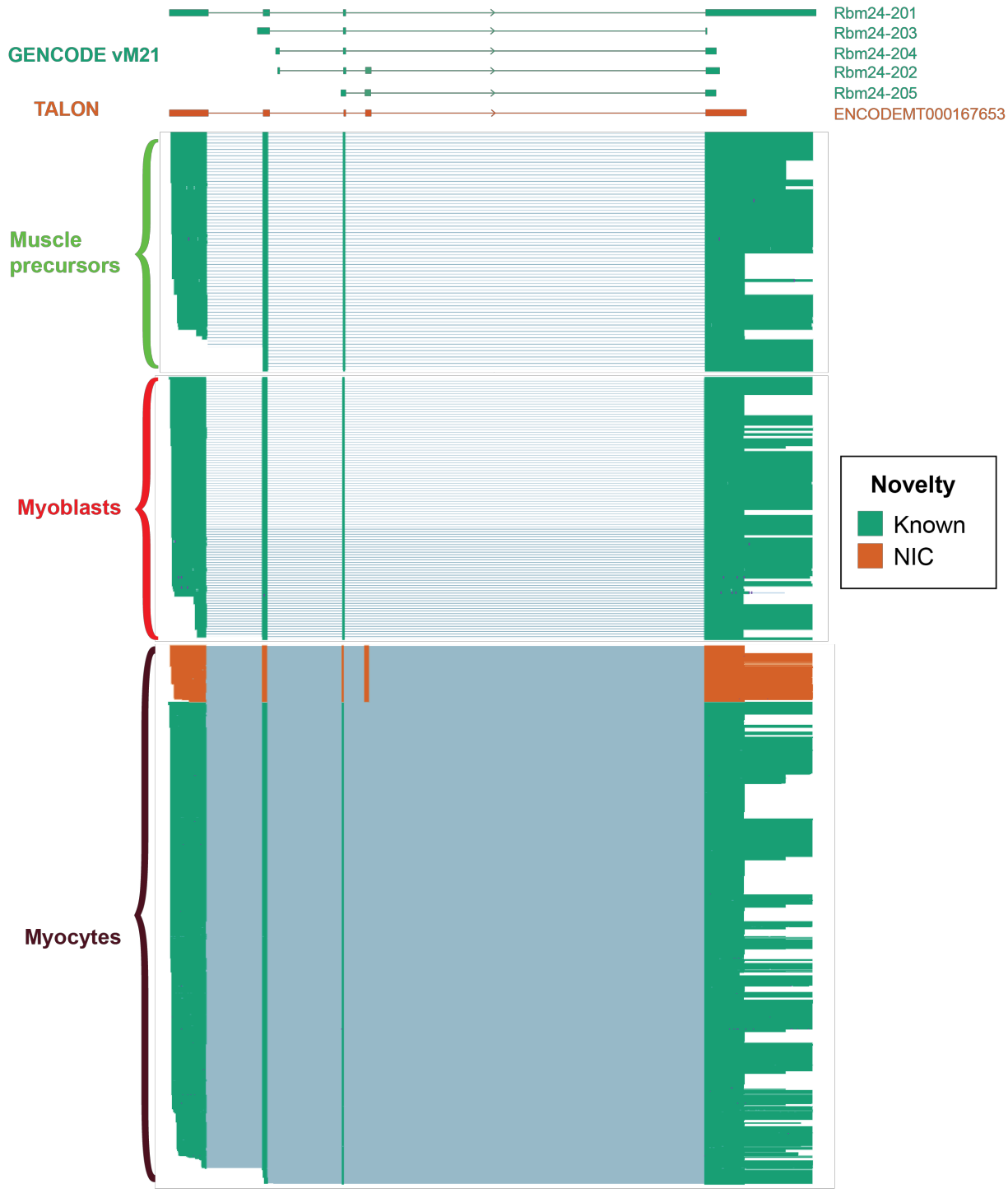

Figure 4.14. A novel Rbm24 isoform is exclusively expressed by myocyte cells. Each line represents an individual long read, and the color identifies the novelty type of the isoform that the read was assigned to. All Rbm24 reads are shown for each cell type. The myocyte track had more Rbm24 reads than the other two and was compressed vertically to allow all reads to be displayed.

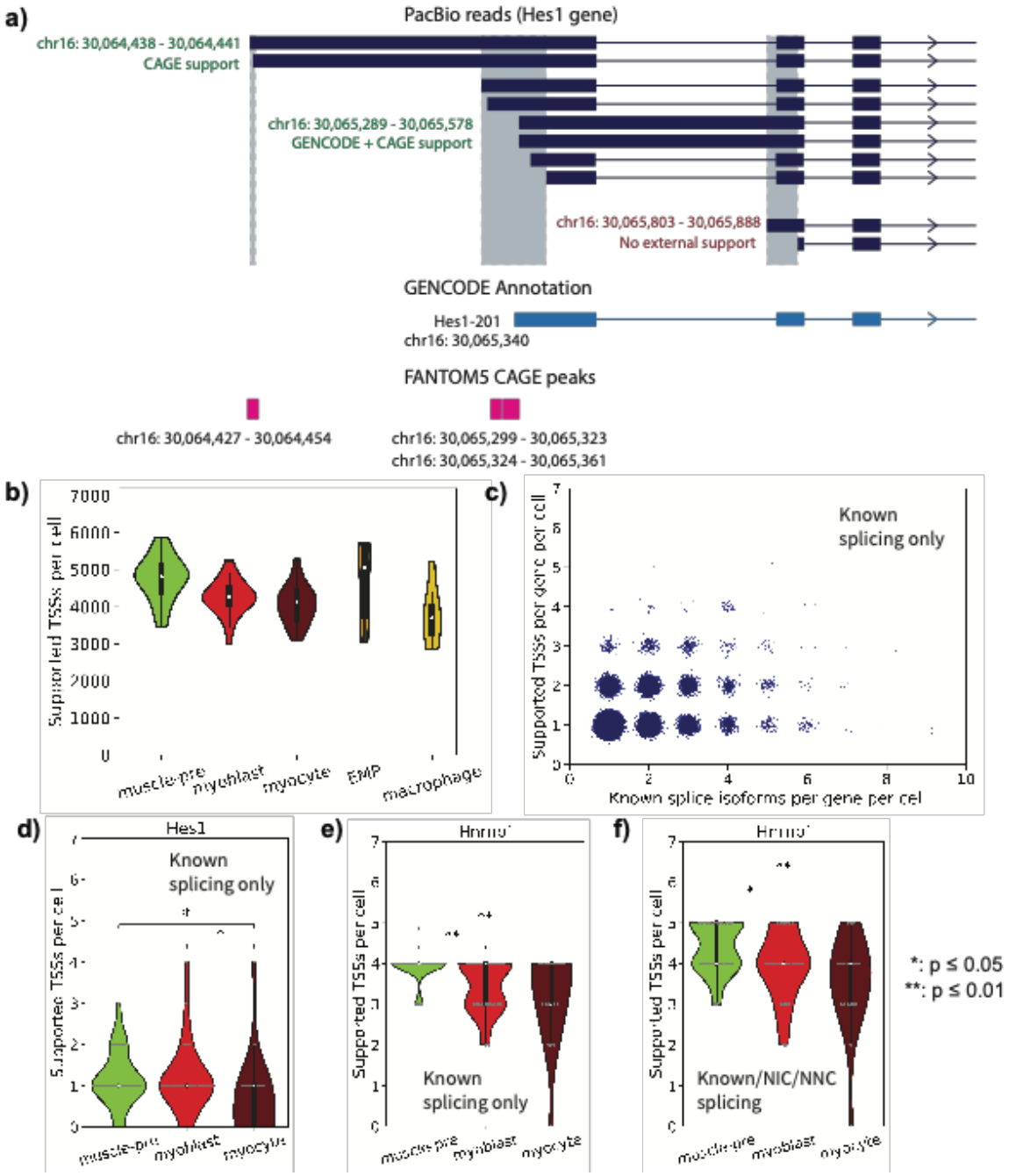

Figure 4.15. Single-cell detection of TSSs in long reads. a) Example of TSS calling in the *Hes1* gene. Consecutive read starts are included in the same TSS if they are ≤ 50 bp apart. A TSS is 'supported' if it has > 1 read and overlaps a GENCODE start site and/or CAGE peak (10 bp flexibility). Note: Subset of reads and TSSs at locus are shown. Scale is approximate. **b)** Distinct TSSs detected per cell by type. **c)** Known splice isoforms detected per gene per cell versus the number of distinct TSSs detected for that same gene in that cell. **d)** Supported *Hes1* TSSs detected in each cell when only reads from known splice isoforms are counted. **e)** Supported *Hnrnpf* TSSs detected in each cell when only reads from known splice isoforms are counted, and **f)** when reads from known/NIC/NNC splice isoforms are counted. Pairwise P-values in panels d, e, and f come from the Mann-Whitney *U* test.

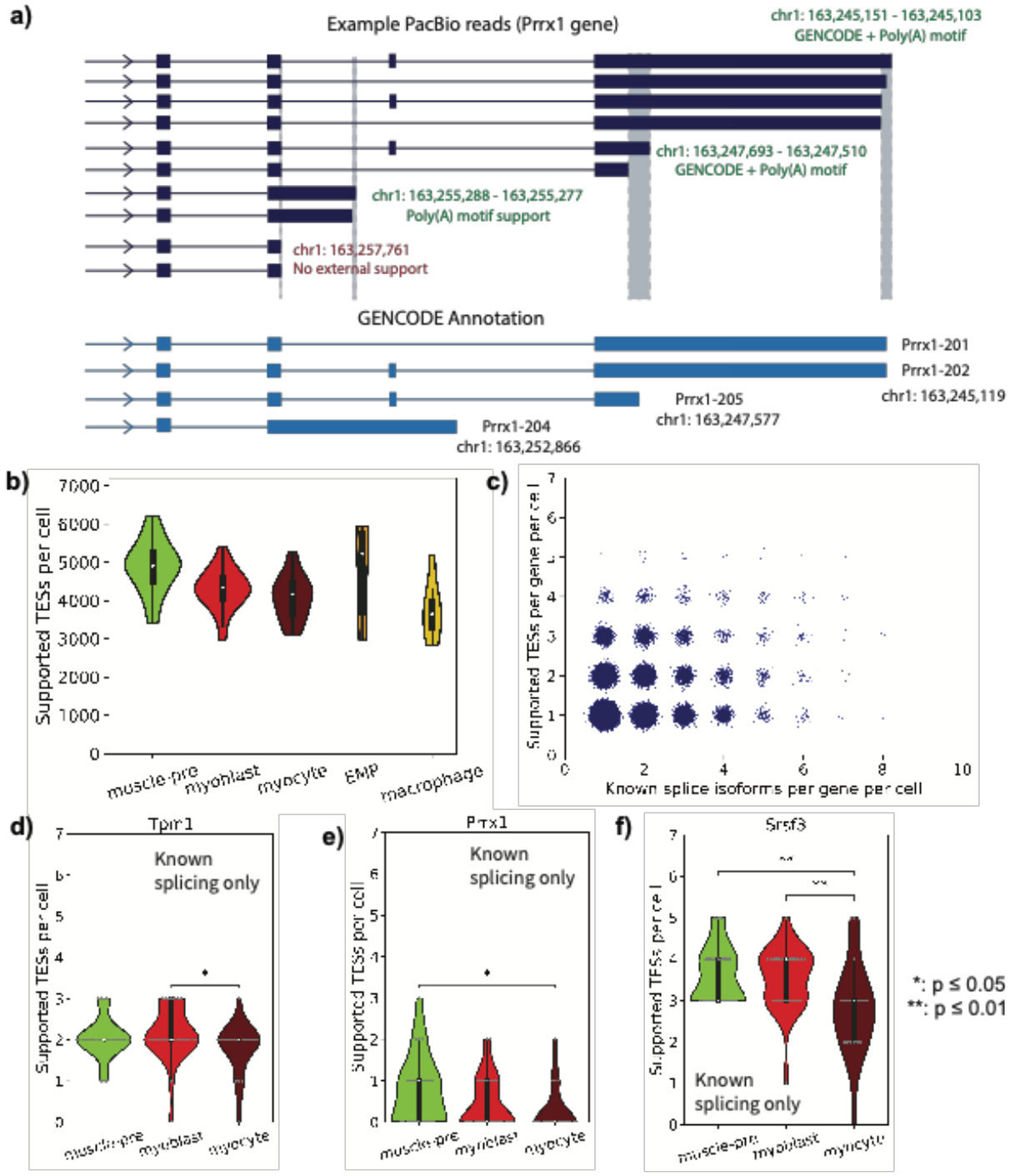

Figure 4.16. Single-cell detection of TESs in long reads. a) Example of TES calling in the *Prrx1* gene. Consecutive read ends are included in the same TES if they are ≤ 100 bp apart. A TES is 'supported' if it has > 1 read and overlaps a GENCODE end site and/or has poly-(A) motif support. Note: Subset of reads and TESs at locus are shown. Scale is approximate. **b)** Distinct TESs detected per cell by type. **c)** Known splice isoforms detected per gene per cell versus the number of distinct TSSs detected for that same gene in that cell. **d-f)** Number of supported TESs detected in each cell from known splice isoforms for the Tpm1, Prrx1, and Srsf3 genes, respectively. Pairwise P-values come from the Mann-Whitney *U* test.

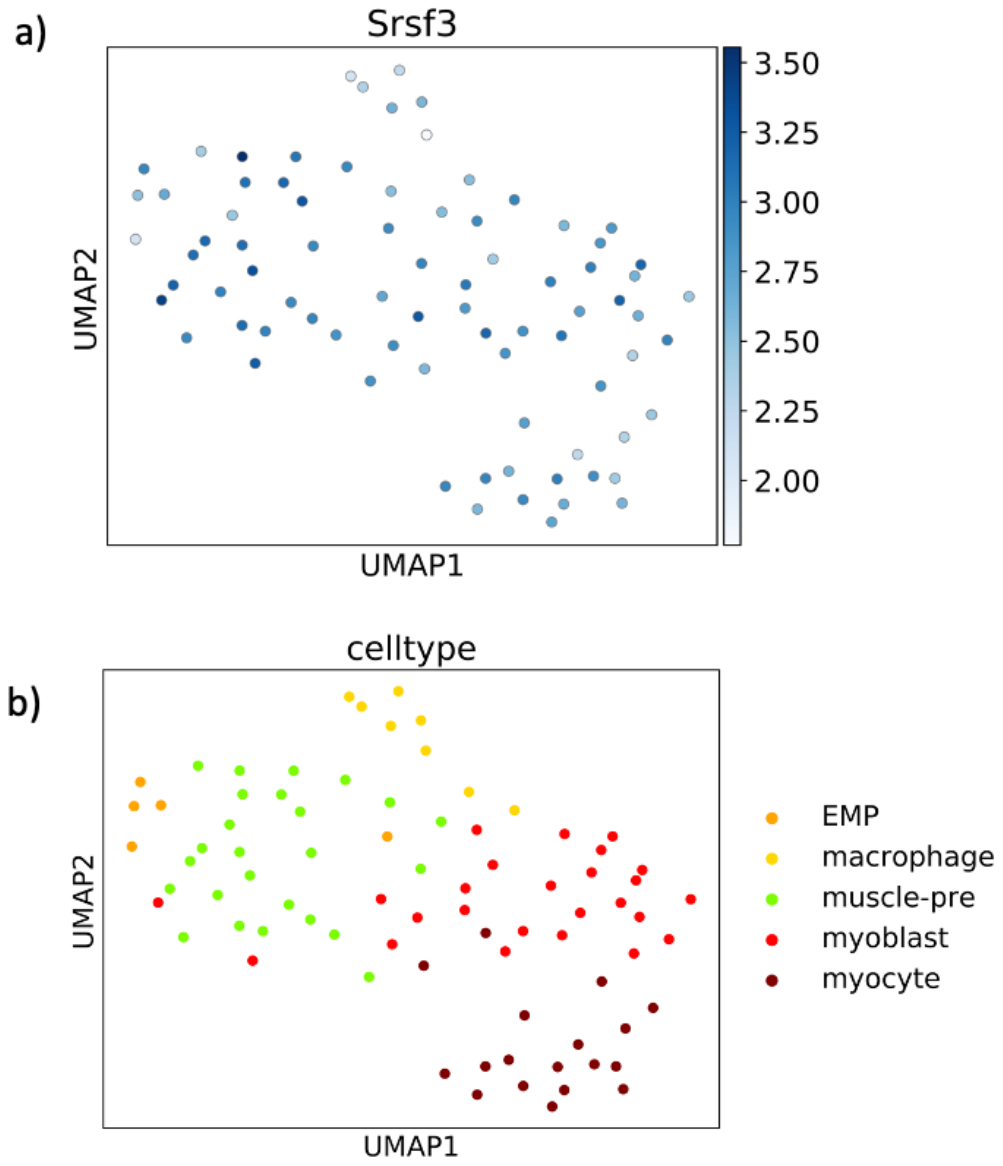

Figure 4.17. Long-read *Srsf3* expression in single cells. a) Expression of Srsf3 is shown on the Scanpy UMAP derived from long-read gene expression. **b)** Celltype labels are shown for reference.

Table 4.1. Scanpy gene markers called for muscle-pre cluster (Leiden 0)

Gene	Score	Log Fold Change	Pval adj.	cluster
Car14	5.4020295	3.0252903	0.0001858	muscle-pre
Gm49708	4.633201	4.2233005	0.00220574	muscle-pre
Hsd11b2	4.5118074	3.8775368	0.00302087	muscle-pre
Pax3	4.5118074	3.5957007	0.00302087	muscle-pre
Nkd2	4.4409943	3.0505488	0.00352325	muscle-pre
Macrod2	4.3954716	2.2377129	0.00398779	muscle-pre
Ints2	4.304426	1.9214656	0.00488374	muscle-pre
Plscr1	4.2083225	2.0139935	0.00630842	muscle-pre
Aif1l	4.1273932	1.8997741	0.00721809	muscle-pre
Tes	4.05658	1.7020516	0.00896284	muscle-pre
Tspan13	4.046464	1.7668782	0.00926042	muscle-pre
E2f5	3.9554183	1.6783144	0.01164527	muscle-pre
Mogat2	3.9402442	4.4591904	0.01229629	muscle-pre
Osr2	3.9351861	1.8555076	0.01233785	muscle-pre
Dhcr24	3.8542569	2.0877926	0.01623129	muscle-pre
Ccnf	3.8036761	1.5469029	0.01773566	muscle-pre
Tcf15	3.8036761	2.6894221	0.01773566	muscle-pre
Depdc1b	3.7834437	2.4781637	0.0185597	muscle-pre
Acot1	3.7733276	2.233109	0.01905867	muscle-pre
Mycn	3.6873403	1.785097	0.02426595	muscle-pre
Hspbap1	3.6721659	1.564011	0.02543173	muscle-pre
Eipr1	3.6620498	1.4385171	0.0262933	muscle-pre

Table 4.2. Scanpy gene markers called for myoblast clusters (Leiden 1 & 3)

Gene	Score	Log Fold Change	Pval adj.	cluster
Vgll3	3.700404	2.2235928	0.19168029	myoblast.1
Msc	3.4663422	2.4202578	0.30299196	myoblast.1
Ogn	3.36603	1.640408	0.32886577	myoblast.1
Nab1	3.315874	2.0490253	0.32886577	myoblast.1
Plagl1	3.2434263	2.0487673	0.34027715	myoblast.1
Col5a1	3.1876974	2.583274	0.35681809	myoblast.1
Deaf1	3.1876974	1.5323653	0.35681809	myoblast.1
Slc7a5	3.1096768	1.5289271	0.40622856	myoblast.1
Ddah1	3.0706666	1.6460644	0.41535473	myoblast.1
Pum2	3.042802	1.6736284	0.43577853	myoblast.1
Mgp	3.042802	1.992388	0.43577853	myoblast.1
Nbl1	2.9870732	1.6416957	0.45821597	myoblast.1
Kcnk1	2.9647815	2.5078528	0.47451216	myoblast.1
Spr	2.9647815	1.5565022	0.47451216	myoblast.1
Plat	2.8979068	2.5100403	0.512567	myoblast.1
Xxylt1	2.8979068	1.5228835	0.512567	myoblast.1
Akt1s1	2.886761	1.6203117	0.51853389	myoblast.1
Zc3hav1l	2.8533235	2.5995398	0.53430141	myoblast.1
Snhg18	2.8533235	1.9643141	0.53430141	myoblast.1
Agtr2	2.8366048	2.1542766	0.53767757	myoblast.1
Lfng	2.8366048	1.5166178	0.53767757	myoblast.1
Susd6	2.8143132	2.8942626	0.54998629	myoblast.1
Sncaip	2.7975945	1.8104236	0.55482138	myoblast.1
Angptl2	2.775303	1.9029424	0.56895612	myoblast.1
Rtl3	2.7307198	1.5311079	0.62527262	myoblast.1
Chodl	2.7251468	3.3846366	0.63195917	myoblast.1
Abcb10	2.719574	1.7912517	0.63195917	myoblast.1
Gpr107	2.719574	1.8943115	0.63195917	myoblast.1
Pus3	2.6972823	1.2959712	0.65750809	myoblast.1
Kdelr3	2.6471262	1.9617811	0.69018488	myoblast.1
Dlg3	2.6415534	2.0723667	0.69122316	myoblast.1
Mmp11	2.6304076	1.4631004	0.69995637	myoblast.1
Osbpl2	2.6304076	1.588369	0.69995637	myoblast.1
Dcx	2.6192617	5.571286	0.69995637	myoblast.1
Moxd1	4.0094733	3.089231	0.19069019	myoblast.2
Vgll3	3.897058	2.4300573	0.19271161	myoblast.2
Msc	3.8845675	2.8909311	0.19271161	myoblast.2
Sapcd2	3.5723033	2.0998673	0.39243569	myoblast.2
Depdc1a	3.5723033	2.1475537	0.39243569	myoblast.2
Fap	3.3349824	2.0625885	0.49768006	myoblast.2
Six2	3.241303	2.035452	0.54408127	myoblast.2
Nid2	3.197586	2.101204	0.58005404	myoblast.2
Fbn2	2.972756	1.8771782	0.84634908	myoblast.2
Cdc25c	2.9477746	1.6738876	0.85544397	myoblast.2
Nek1	2.8416047	1.6602198	0.91499369	myoblast.2
Ebf3	2.8041332	1.9811496	0.9778417	myoblast.2
Nxt2	2.7978878	1.6335794	0.9778417	myoblast.2
Ptpn9	2.6917179	1.5563906	0.98827117	myoblast.2
Dmrt2	2.6604915	1.5260006	0.98827117	myoblast.2

Table 4.3. Scanpy gene markers called for myocyte cluster (Leiden 2)

Gene	Score	Log Fold	Pval adj.	cluster
Mymk	6.1683693	9.132295	2.25E-06	myocyte
Actc1	6.1683693	7.3800144	2.25E-06	myocyte
Zbtb18	6.1683693	6.9418387	2.25E-06	myocyte
Fndc5	6.1683693	4.702166	2.25E-06	myocyte
Myog	6.1683693	7.7061925	2.25E-06	myocyte
Mymx	6.12092	7.453832	2.25E-06	myocyte
Atp2a1	6.049747	5.033737	3.07E-06	myocyte
Iffo1	6.0260224	5.852096	3.16E-06	myocyte
Cap2	5.966711	4.3768435	3.41E-06	myocyte
Cryab	5.9785733	5.2330327	3.41E-06	myocyte
Rbm24	5.931124	4.1954484	3.92E-06	myocyte
Arpp21	5.895538	5.303498	3.95E-06	myocyte
Klhl41	5.7472596	8.508643	8.08E-06	myocyte
Kremen2	5.6642237	6.2610507	1.03E-05	myocyte
Rgs16	5.670155	3.8824253	1.03E-05	myocyte
Ttn	5.658293	4.5299516	1.03E-05	myocyte
Chrna1	5.670155	6.1346893	1.03E-05	myocyte
Tnni1	5.628637	6.4922442	1.14E-05	myocyte
Neb	5.5989814	4.1797447	1.26E-05	myocyte
4-Sep	5.397323	30.253965	3.37E-05	myocyte
Rb1	5.397323	4.0720644	3.37E-05	myocyte
Nes	5.3676677	3.8621235	3.86E-05	myocyte
Gadd45g	5.3380117	3.9702566	4.42E-05	myocyte
Tnnt2	5.3142877	6.9242187	4.77E-05	myocyte
Kcnk13	5.2668386	5.80958	5.73E-05	myocyte
Gatm	5.243114	3.3504105	6.36E-05	myocyte
Chrng	5.100767	4.5871997	0.00013008	myocyte
Chrnb1	5.088905	3.8613286	0.0001354	myocyte
Scrib	5.053318	4.261105	0.0001597	myocyte
Cd82	5.017731	3.3554664	0.00018822	myocyte
Plpp1	5.005869	2.4368443	0.00019202	myocyte
Mylpf	5.005869	4.8316493	0.00019202	myocyte
Dll1	4.9821444	4.3206162	0.00021279	myocyte
Lrrn1	4.9465575	2.8981788	0.00025059	myocyte
Vgll2	4.9287643	3.094633	0.00026416	myocyte
Tmem35a	4.910971	3.905218	0.00028395	myocyte
Cdh15	4.8753843	3.2382863	0.00033406	myocyte
Fbxo17	4.8042107	2.5803235	0.0004532	myocyte
Olfml2b	4.780486	4.1567974	0.00050147	myocyte
Trim55	4.7567616	5.499389	0.00055473	myocyte
Hdac11	4.7330375	4.4299636	0.000594	myocyte
Des	4.715244	3.1071703	0.00060946	myocyte
Cdkn1a	4.6974506	3.2640507	0.00063671	myocyte
Dync1i1	4.679657	4.680893	0.00068475	myocyte
Fitm1	4.6381392	2.1441612	0.0008148	myocyte
S100a16	4.6025524	7.855525	0.0009542	myocyte
Crip2	4.59069	3.652356	0.00098349	myocyte
Traf3ip3	4.59069	3.4632251	0.00098349	myocyte
Hfe2	4.5551033	6.987626	0.00115049	myocyte

Tmem200a	4.519517	6.310513	0.00132741	myocyte
Lmna	4.4779987	2.565389	0.00155461	myocyte
Dner	4.4246187	4.84567	0.00185749	myocyte
Adgre5	4.4246187	2.9423041	0.00185749	myocyte
Myzap	4.400894	4.7150245	0.0020494	myocyte
Hspb1	4.3653073	2.9702024	0.00230943	myocyte
Ccdc141	4.329721	3.6239035	0.00268721	myocyte
Map4	4.3059964	3.3798952	0.00292973	myocyte
Palmd	4.3059964	3.8218606	0.00292973	myocyte
Pip4k2a	4.300065	2.2182345	0.00297823	myocyte
Rtn4	4.282272	2.615745	0.00312993	myocyte
Myl1	4.2763405	4.0295606	0.00318262	myocyte
Ablim3	4.240754	29.484991	0.00362409	myocyte

Table 4.4. Scanpy gene markers called for macrophage cluster (Leiden 4)

Gene	Score	Log Fold Change	Pval adj.	cluster
Nfam1	4.057513	31.389265	0.0155388	macrophage
Tapbp	3.9673464	5.2011533	0.0155388	macrophage
Rtp4	3.9673464	5.487895	0.0155388	macrophage
Tifa	3.9853797	6.211637	0.0155388	macrophage
Lcp1	3.9853797	6.408189	0.0155388	macrophage
Slc7a7	3.9853797	6.9965487	0.0155388	macrophage
Fcer1g	4.003413	5.6955624	0.0155388	macrophage
Casp1	4.003413	6.6441946	0.0155388	macrophage
Tnfaip8l2	4.003413	6.4254136	0.0155388	macrophage
Trf	4.003413	7.6763115	0.0155388	macrophage
Cd68	4.003413	6.659149	0.0155388	macrophage
Tmem86a	4.003413	4.0100503	0.0155388	macrophage
Maf	4.003413	5.9239583	0.0155388	macrophage
Lyn	3.9673464	5.730999	0.0155388	macrophage
Ptpn6	3.9312797	5.9449954	0.0155388	macrophage
Ostf1	3.9312797	4.3841844	0.0155388	macrophage
Pf4	3.9312797	5.0301304	0.0155388	macrophage
Pld4	3.9312797	6.7817526	0.0155388	macrophage
Pon3	3.9312797	5.051027	0.0155388	macrophage
Tyrobp	3.9312797	5.3925257	0.0155388	macrophage
Psmb9	3.949313	5.414737	0.0155388	macrophage
Gpr34	3.949313	5.886734	0.0155388	macrophage
Hexb	3.949313	4.2991667	0.0155388	macrophage
Ctsc	3.949313	6.0997887	0.0155388	macrophage
Tmem37	3.949313	5.7186313	0.0155388	macrophage
Il4ra	3.949313	5.354958	0.0155388	macrophage
Unc93b1	4.003413	6.7391615	0.0155388	macrophage
Anxa3	4.0214467	5.7330155	0.0155388	macrophage
Slfn2	4.057513	7.5114217	0.0155388	macrophage
Hpgds	4.0214467	7.38303	0.0155388	macrophage
Phyhd1	4.057513	8.150206	0.0155388	macrophage
Spi1	4.057513	7.073855	0.0155388	macrophage
P2ry12	4.057513	10.195053	0.0155388	macrophage
Csf1r	4.057513	12.443872	0.0155388	macrophage
Tmem106a	4.057513	7.44141	0.0155388	macrophage
Trem2	4.057513	11.917926	0.0155388	macrophage
Aif1	4.057513	30.64478	0.0155388	macrophage
C1qa	4.057513	13.540819	0.0155388	macrophage
C1qb	4.057513	13.081396	0.0155388	macrophage
C1qc	4.057513	12.657262	0.0155388	macrophage
Lrrc25	4.057513	28.026665	0.0155388	macrophage
C3ar1	4.057513	12.056113	0.0155388	macrophage
C5ar1	4.057513	30.908575	0.0155388	macrophage
Ms4a6d	4.057513	31.657768	0.0155388	macrophage
Gpr160	4.057513	30.793812	0.0155388	macrophage
Cd86	4.057513	11.483597	0.0155388	macrophage
Fcrls	4.057513	13.164282	0.0155388	macrophage

Mrc1	4.057513	32.394344	0.0155388	macrophage
Ms4a6b	4.0214467	7.577761	0.0155388	macrophage
Cx3cr1	4.0214467	8.526653	0.0155388	macrophage
Il10rb	4.0214467	5.6306634	0.0155388	macrophage
Apoe	4.057513	8.440612	0.0155388	macrophage
Ccl9	4.0214467	7.156742	0.0155388	macrophage
Irf8	4.0394797	6.5784955	0.0155388	macrophage
Irf5	4.0394797	7.066024	0.0155388	macrophage
Sirpa	4.0394797	6.93287	0.0155388	macrophage
Ifi204	4.0214467	6.6691546	0.0155388	macrophage
Adgre1	4.0394797	7.978259	0.0155388	macrophage
Clec4a2	4.057513	9.093093	0.0155388	macrophage
Ctss	4.0394797	7.4226303	0.0155388	macrophage
Uap1l1	3.8771794	4.9545493	0.0161774	macrophage
Cfh	3.9132462	5.833235	0.0161774	macrophage
Ifi30	3.9132462	5.272209	0.0161774	macrophage
Klf2	3.895213	3.8315318	0.0161774	macrophage
Arhgap30	3.895213	5.0778165	0.0161774	macrophage

Table 4.5. Scanpy isoform markers called for muscle precursor clusters (Leiden 1,3)

Isoform	Novelty	Score	Log Fold Change	Pval adj.	Cluster
Kif18a-201	Known	4.8418484	2.5332475	0.00292975	muscle-pre.1
Depdc1b-201	Known	4.7250495	3.1452296	0.00425725	muscle-pre.1
Ccnf-201	Known	4.650723	2.4526238	0.00544086	muscle-pre.1
Bora-201	Known	4.459597	2.0784216	0.0097249	muscle-pre.1
Knl1-201	Known	4.337489	2.0168455	0.01410105	muscle-pre.1
Depdc1a-203	Known	4.3215623	2.8921287	0.01433475	muscle-pre.1
Sapcd2-201	Known	4.087964	2.8468535	0.03458265	muscle-pre.1
Gm10184-201	Known	4.077346	2.0365045	0.03458265	muscle-pre.1
Prc1-206	Known	4.077346	1.9713358	0.03458265	muscle-pre.1
Nuf2-201	Known	4.0189466	1.7332197	0.03933754	muscle-pre.1
Ska3-201	Known	3.9977105	1.7515329	0.04120397	muscle-pre.1
Dbf4-203	Known	3.9924014	1.7725917	0.04120397	muscle-pre.1
Ttk-201	Known	3.9711652	1.7023541	0.04411713	muscle-pre.1
Nusap1-201	Known	3.8543663	1.8275551	0.06481907	muscle-pre.1
Oip5-202	Known	3.8331301	1.658908	0.06689453	muscle-pre.1
Cdca5-201	Known	3.7694216	1.6803302	0.08074415	muscle-pre.1
Ndc80-201	Known	3.7534943	1.5409368	0.08327797	muscle-pre.1
Ncapg-201	Known	3.6632407	1.7917058	0.10843686	muscle-pre.1
Knstrn-206	Known	3.6526225	1.5931767	0.1113843	muscle-pre.1
Entr1-201	Known	3.6366954	2.4724135	0.11356238	muscle-pre.1
Coasy-201	Known	3.6207683	1.5757657	0.11752356	muscle-pre.1
Kif11-201	Known	3.578296	1.4158559	0.12799604	muscle-pre.1
Poc1a-201	Known	3.5570598	1.6891567	0.1306336	muscle-pre.1
Cip2a-201	Known	3.5570598	1.7929885	0.1306336	muscle-pre.1
Hsd11b2-201	Known	3.5464416	2.712209	0.13443107	muscle-pre.1
Prc1-202	Known	3.5411327	1.7184653	0.13558875	muscle-pre.1
Ckap2-201	Known	3.4243336	1.612998	0.18065494	muscle-pre.1
Kpna2-ENCODEMT000383037	NIC	3.4243336	2.1289346	0.18065494	muscle-pre.1
Macrod2-204	Known	3.4084065	1.9190261	0.18425616	muscle-pre.1
Terf1-203	Known	3.3977883	1.4129584	0.18975081	muscle-pre.1
Cenph-201	Known	3.3871703	1.2960953	0.19540598	muscle-pre.1
Nek2-201	Known	3.3712432	2.112818	0.20514444	muscle-pre.1
Ube2t-201	Known	3.3128438	1.4995973	0.23471287	muscle-pre.1
Cdca2-204	Known	3.3022256	2.015715	0.23471287	muscle-pre.1
Tpx2-202	Known	3.3022256	2.217055	0.23471287	muscle-pre.1
Cenpl-203	Known	3.2916076	1.6752703	0.2359557	muscle-pre.1
Psrc1-201	Known	3.2916076	1.8760694	0.2359557	muscle-pre.1
Daxx-201	Known	3.2703714	1.6053554	0.25040066	muscle-pre.1
Cdc25c-201	Known	3.206663	1.3763092	0.29893906	muscle-pre.1
Mbip-201	Known	3.2013538	1.4214157	0.29999152	muscle-pre.1
Gemin6-ENCODEMT000278446	NNC	3.1854267	1.5211272	0.31011324	muscle-pre.1
Bub1b-201	Known	3.1854267	1.5196954	0.31011324	muscle-pre.1
Osr2-201	Known	4.046945	2.9633505	0.54014211	muscle-pre.2
Ptma-ENCODEMT000223015	NIC	2.5917933	1.2528777	0.98456262	muscle-pre.2
Dcp2-201	Known	2.62302	1.5583023	0.98456262	muscle-pre.2
Tmem30b-201	Known	2.629265	2.4316359	0.98456262	muscle-pre.2
Slc25a5-ENCODEMT000161559	NNC	2.6729822	1.4240137	0.98456262	muscle-pre.2

Tpm1-213	Known	2.6854727	2.0807443	0.98456262	muscle-pre.2
Ldhb-ENCODEMT000433208	NIC	2.5730577	1.3159151	0.98456262	muscle-pre.2
Ndufaf5-201	Known	2.5480764	1.3475776	0.98456262	muscle-pre.2
B9d2-203	Known	2.4543972	1.2891717	0.98456262	muscle-pre.2
Nicn1-201	Known	2.4606423	1.1927199	0.98456262	muscle-pre.2
Nkd2-201	Known	2.4606423	1.687973	0.98456262	muscle-pre.2
Ddt-201	Known	2.5106046	1.6233656	0.98456262	muscle-pre.2
Rrm2-ENCODEMT000436459	NNC	2.5230954	1.7915854	0.98456262	muscle-pre.2
Mogat2-201	Known	2.5230954	2.0340626	0.98456262	muscle-pre.2
Car14-201	Known	2.6979632	1.6422741	0.98456262	muscle-pre.2
Srsf7-205	Known	3.085171	2.4485793	0.98456262	muscle-pre.2
Hnrnpc-ENCODEMT000273062	NIC	3.2350578	2.2577627	0.98456262	muscle-pre.2
Tipin-205	Known	3.322492	1.6066203	0.98456262	muscle-pre.2
Mpg-201	Known	2.997737	2.0498493	0.98456262	muscle-pre.2
Riox1-201	Known	2.7042086	1.5094658	0.98456262	muscle-pre.2
Tes-202	Known	2.9914916	1.8418247	0.98456262	muscle-pre.2
Trappc1-202	Known	2.822869	1.5575259	0.98456262	muscle-pre.2
Rpl15-203	Known	2.8353596	1.4044408	0.98456262	muscle-pre.2
Mcm5-205	Known	2.9602652	1.7092898	0.98456262	muscle-pre.2
Cnih2-201	Known	2.4356613	1.725709	0.98456262	muscle-pre.2

Table 4.6. Scanpy isoform markers called for myoblast cluster (Leiden 0)

Isoform	Novelty	Score	Log Fold Change	Pval adj.	Cluster
Msc-201	Known	6.0816016	4.459252	1.76E-05	myoblast
Vgll3-202	Known	5.7868304	4.4114966	5.48E-05	myoblast
Slc38a4-201	Known	5.233487	2.4577858	0.00054723	myoblast
Klhl13-ENCODEMT000158660	NIC	4.7577157	2.22412	0.00483094	myoblast
Spry1-ENCODEMT000164274	NIC	4.5198298	2.4397829	0.01077894	myoblast
Fxyd6-201	Known	4.219887	1.7706419	0.03446158	myoblast
Rtl3-201	Known	4.1164584	2.238378	0.05177886	myoblast
Moxd1-201	Known	4.0699153	3.1572297	0.06054219	myoblast
Chodl-ENCODEMT000326138	NNC	4.002687	6.0318184	0.07417062	myoblast
Spry1-201	Known	3.992344	2.3537247	0.07450213	myoblast
Msc-204	Known	3.9613154	3.609085	0.07881861	myoblast
Myf5-201	Known	3.8734012	2.027822	0.10957714	myoblast
Cited1-202	Known	3.837201	2.4036095	0.1228185	myoblast
Lfng-201	Known	3.7544582	1.8220166	0.14287244	myoblast
Shisal2a-201	Known	3.7130868	3.7927008	0.1638416	myoblast
Arhgap29-ENCODEMT000263463	NIC	3.6872296	2.3747003	0.17211266	myoblast
Fgfr4-201	Known	3.6562011	3.2005439	0.18045416	myoblast
Gart-202	Known	3.6044867	1.6802435	0.20578904	myoblast
Entpd4-206	Known	3.5372581	1.9686908	0.25469451	myoblast
Prkci-201	Known	3.5217438	1.4278562	0.25905077	myoblast
Nfib-202	Known	3.464858	1.5686892	0.2855887	myoblast
Dlk1-208	Known	3.4700296	1.7494532	0.2855887	myoblast
Arfip2-206	Known	3.464858	1.4423202	0.2855887	myoblast
Pitx3-201	Known	3.4700296	1.4916306	0.2855887	myoblast
Chodl-201	Known	3.4855437	5.04625	0.2855887	myoblast
Tmem246-201	Known	3.4338295	2.5512786	0.31464756	myoblast
P3h2-201	Known	3.4183152	3.4937754	0.32729867	myoblast
Top2b-201	Known	3.402801	1.6115663	0.3291366	myoblast
Fzd4-201	Known	3.3924582	3.6997597	0.33621353	myoblast
Mta1-204	Known	3.3355725	1.7063799	0.37617183	myoblast
Agtr2-201	Known	3.3097153	2.495194	0.39501699	myoblast
Entpd4b-204	Known	3.2786865	2.0241003	0.43490972	myoblast
Msc-ENCODEMT000144365	NNC	3.2166295	4.459759	0.509456	myoblast
Cnot10-201	Known	3.2062867	1.5141709	0.509456	myoblast
Ddx51-201	Known	3.139058	2.1518767	0.57388791	myoblast
Pdgfa-202	Known	3.1338866	1.3962227	0.57388791	myoblast
Gm266-201	Known	3.1235437	3.5360732	0.58784153	myoblast
Mgp-201	Known	3.102858	1.4088854	0.5973385	myoblast
Rps8-203	Known	2.9632294	1.5586787	0.81203166	myoblast
Zfp580-201	Known	2.9270294	1.3217518	0.84436227	myoblast
Parm1-201	Known	2.8960009	1.7470729	0.90978745	myoblast

Table 4.7. Scanpy isoform markers called for myocyte cluster (Leiden 2)

Isoform	Novelty	Score	Log Fold Change	Pval adj.	Cluster
Actc1-201	Known	6.018265	7.3777175	1.76E-05	myocyte
Myog-201	Known	6.006107	7.7137384	1.76E-05	myocyte
Mymk-201	Known	5.8723674	6.580818	2.12E-05	myocyte
Fndc5-201	Known	5.702154	3.723944	4.38E-05	myocyte
Rgs16-201	Known	5.6535215	4.59659	4.47E-05	myocyte
Rbm24-201	Known	5.6413636	3.9933963	4.47E-05	myocyte
Tnnt1-203	Known	5.629205	3.4453943	4.47E-05	myocyte
Klhl41-201	Known	5.580573	8.610352	4.76E-05	myocyte
Zbtb18-202	Known	5.5684147	7.8576965	4.76E-05	myocyte
Mymx-203	Known	5.5684147	6.7762823	4.76E-05	myocyte
Bin1-206	Known	5.5319405	3.338935	5.21E-05	myocyte
Atp2a1-ENCODEMT000341384	NIC	5.507624	4.978309	5.49E-05	myocyte
Tnni1-ENCODEMT000266900	NIC	5.495466	6.713701	5.49E-05	myocyte
Chrna1-201	Known	5.4650707	6.0732727	6.23E-05	myocyte
Cap2-201	Known	5.4468336	4.7288866	6.55E-05	myocyte
Cryab-201	Known	5.4407544	5.4953322	6.55E-05	myocyte
Slc25a3-206	Known	5.3921223	5.3179917	7.36E-05	myocyte
Eno1-ENCODEMT000412975	NIC	5.40428	5.4388027	7.36E-05	myocyte
Tnnt1-ENCODEMT000153433	NIC	5.300936	2.7793067	0.0001066	myocyte
Tmed2-ENCODEMT000301604	NIC	5.167197	7.7333364	0.00018515	myocyte
Nes-201	Known	5.161118	3.8478608	0.00018636	myocyte
Lsm6-204	Known	5.1064067	3.917087	0.00022094	myocyte
Gatm-201	Known	5.1185646	3.3676765	0.00022094	myocyte
Gadd45g-201	Known	5.1124854	3.944901	0.00022094	myocyte
Lrrn1-201	Known	5.1064067	3.2179692	0.00022094	myocyte
Mylpf-201	Known	5.0699325	6.146232	0.00026184	myocyte
Cdkn1c-ENCODEMT000373118	NIC	5.045616	3.342858	0.00028478	myocyte
Hacd1-203	Known	5.0213	5.6134887	0.00031012	myocyte
Myl1-201	Known	5.0213	6.1735725	0.00031012	myocyte
Neb-204	Known	5.0091414	4.110547	0.00032375	myocyte
Rb1-201	Known	4.948351	4.2429376	0.00042615	myocyte
Vgll2-202	Known	4.942272	3.260935	0.00043137	myocyte
Zbtb18-201	Known	4.8997188	4.747299	0.00050743	myocyte
Cdh15-201	Known	4.8875604	3.5815728	0.00053031	myocyte
Scrib-ENCODEMT000187640	NIC	4.814612	29.84949	0.00074013	myocyte
Serinc2-203	Known	4.7902956	4.1015873	0.0008218	myocyte
Vgll2-201	Known	4.7538214	2.9466174	0.00095322	myocyte
Des-201	Known	4.723426	3.3049855	0.00108988	myocyte
Actc1-ENCODEMT000343314	NNC	4.6869516	5.1047854	0.00126314	myocyte
Mymx-ENCODEMT000242249	NNC	4.6747937	5.4420204	0.0012812	myocyte
Actc1-ENCODEMT000342824	NIC	4.6747937	5.498227	0.0012812	myocyte
Tnnt1-206	Known	4.6626353	2.8404799	0.00133959	myocyte
Cdkn1a-201	Known	4.644398	3.3985753	0.00140255	myocyte
Vasp-201	Known	4.644398	3.4756563	0.00140255	myocyte
Cdkn1c-ENCODEMT000373182	NIC	4.620082	3.1417873	0.0015143	myocyte
Tnnt1-201	Known	4.614003	3.05506	0.00153878	myocyte
Cdkn1c-ENCODEMT000373034	NIC	4.510659	2.4062722	0.0024221	myocyte

Chrng-201	Known	4.468106	4.667389	0.00292066	myocyte
Crip2-201	Known	4.4255524	3.6297178	0.00339067	myocyte
Mymx-201	Known	4.3769197	6.761977	0.00404984	myocyte
Shisa2-201	Known	4.3586826	2.6834664	0.00435296	myocyte
Lmna-203	Known	4.352604	2.8947704	0.00437711	myocyte
Vma21-203	Known	4.3282876	6.450808	0.00483619	myocyte
Pnmal2-201	Known	4.3222084	3.6279802	0.00491796	myocyte
Cd82-201	Known	4.316129	3.174818	0.00494887	myocyte
Tmem35a-201	Known	4.316129	3.6898923	0.00494887	myocyte
Acta2-ENCODEMT000198548	NNC	4.273576	3.1073308	0.00587162	myocyte
Ube2d3-ENCODEMT000273941	NIC	4.267497	5.277565	0.00597231	myocyte
Chrnb1-201	Known	4.261418	3.7903194	0.00607514	myocyte
Palmd-201	Known	4.2431808	4.028244	0.00652483	myocyte
Lsm6-ENCODEMT000410310	NNC	4.2310224	4.984635	0.00681955	myocyte
Vdac3-ENCODEMT000393107	NNC	4.2127852	3.1398304	0.00725082	myocyte
Hspb1-201	Known	4.2067065	3.0288143	0.00737693	myocyte
Trim55-202	Known	4.194548	4.758986	0.00763707	myocyte

Table 4.8. Scanpy isoform markers called for macrophage cluster (Leiden 4)

Isoform	Novelty	Score	Log Fold Change	Pval adj.	Cluster
C5ar1-202	Known	4.057513	31.207962	0.03270791	macrophage
Anxa3-201	Known	3.9312797	5.87416	0.03270791	macrophage
Pf4-201	Known	3.9312797	5.128383	0.03270791	macrophage
Ifi30-202	Known	3.949313	5.306476	0.03270791	macrophage
Lgmn-ENCODEMT000443455	NNC	3.949313	5.6585793	0.03270791	macrophage
Ctsc-201	Known	3.949313	6.203977	0.03270791	macrophage
Tmem37-201	Known	3.949313	5.727334	0.03270791	macrophage
Tyrobp-202	Known	3.9673464	5.484709	0.03270791	macrophage
Sirpa-ENCODEMT000368623	NIC	3.9673464	6.6034083	0.03270791	macrophage
Rtp4-201	Known	3.9673464	6.037283	0.03270791	macrophage
Fcer1g-201	Known	3.9673464	5.7990203	0.03270791	macrophage
Casp1-201	Known	3.9673464	6.1672387	0.03270791	macrophage
Hexb-201	Known	3.9853797	4.4007444	0.03270791	macrophage
Tmem86a-201	Known	3.9853797	4.311908	0.03270791	macrophage
Pld4-201	Known	3.9312797	6.9321337	0.03270791	macrophage
Mpp1-201	Known	3.9312797	4.1905756	0.03270791	macrophage
Lcp2-201	Known	3.859146	5.2996893	0.03270791	macrophage
Cebpd-201	Known	3.859146	3.9262567	0.03270791	macrophage
Atp6v0a1-ENCODEMT000372261	NIC	3.859146	4.9282146	0.03270791	macrophage
Zfp36-202	Known	3.859146	4.560254	0.03270791	macrophage
Mt1-201	Known	3.859146	4.929718	0.03270791	macrophage
Cd53-201	Known	3.8771794	5.9794636	0.03270791	macrophage
Ctsc-203	Known	3.8771794	5.8942194	0.03270791	macrophage
Maf-202	Known	3.895213	5.3753223	0.03270791	macrophage
Fam49b-207	Known	3.895213	4.1351047	0.03270791	macrophage
Zfp36-201	Known	3.895213	5.6155725	0.03270791	macrophage
Klf2-201	Known	3.895213	3.9960942	0.03270791	macrophage
Psmb10-201	Known	3.9132462	4.556701	0.03270791	macrophage
Ostf1-201	Known	3.9132462	4.4198847	0.03270791	macrophage
Psmb9-204	Known	3.9853797	6.3347144	0.03270791	macrophage
Chd9-ENCODEMT000418033	NNC	4.003413	6.0899177	0.03270791	macrophage
Nfam1-204	Known	4.057513	31.487928	0.03270791	macrophage
Ms4a6d-201	Known	4.057513	31.944609	0.03270791	macrophage
Ctss-202	Known	4.057513	7.818227	0.03270791	macrophage
Dab2-203	Known	4.057513	8.815356	0.03270791	macrophage
Lyn-202	Known	4.057513	7.150842	0.03270791	macrophage
Il10rb-201	Known	4.057513	6.3188634	0.03270791	macrophage
Sirpa-214	Known	4.057513	10.861557	0.03270791	macrophage
Lgmn-202	Known	4.057513	6.673731	0.03270791	macrophage
C1qb-ENCODEMT000401026	NNC	4.057513	30.418652	0.03270791	macrophage
Csf1r-202	Known	4.057513	13.428986	0.03270791	macrophage
Aif1-203	Known	4.057513	30.8233	0.03270791	macrophage
Apoe-206	Known	4.057513	8.509587	0.03270791	macrophage
Spi1-201	Known	4.057513	7.1644115	0.03270791	macrophage
Ifi204-201	Known	4.057513	9.737157	0.03270791	macrophage
C1qa-201	Known	4.057513	33.28197	0.03270791	macrophage
C1qc-201	Known	4.057513	12.49582	0.03270791	macrophage

C1qb-201	Known	4.057513	14.284418	0.03270791	macrophage
Slfn2-201	Known	4.057513	7.542062	0.03270791	macrophage
C3ar1-201	Known	4.057513	11.933406	0.03270791	macrophage
Fyb-210	Known	4.057513	29.61367	0.03270791	macrophage
Fcrls-201	Known	4.057513	12.92575	0.03270791	macrophage
Ifngr1-201	Known	4.0394797	4.342029	0.03270791	macrophage
Cd68-201	Known	4.003413	6.8047543	0.03270791	macrophage
Ptpn6-202	Known	4.003413	6.4402275	0.03270791	macrophage
Sesn1-201	Known	4.003413	4.205529	0.03270791	macrophage
Trf-201	Known	4.003413	7.6765733	0.03270791	macrophage
Grn-201	Known	4.003413	5.042139	0.03270791	macrophage
Irf8-201	Known	4.0214467	7.3081236	0.03270791	macrophage
Ccl9-201	Known	4.0214467	7.219744	0.03270791	macrophage
Ms4a6b-201	Known	4.0214467	7.437858	0.03270791	macrophage
Dab2-202	Known	3.859146	4.1632442	0.03270791	macrophage
Tifa-201	Known	4.0214467	7.034627	0.03270791	macrophage
Unc93b1-203	Known	4.0214467	7.5554075	0.03270791	macrophage
Cndp2-202	Known	4.0214467	4.1234236	0.03270791	macrophage
Tmem106a-201	Known	4.0394797	7.952365	0.03270791	macrophage
Lcp1-ENCODEMT000305365	NNC	4.0394797	7.5133996	0.03270791	macrophage
Hpgds-201	Known	4.0394797	8.171628	0.03270791	macrophage
Cx3cr1-201	Known	4.0214467	8.520628	0.03270791	macrophage

Table 4.9. Isoform-level markers belonging to genes that were not called in the gene-level clustering analysis

Isoform	Score	Log Fold	Pval adj.	Cluster
Kif18a-201	4.8418484	2.5332475	0.00292975	muscle-pre.1
Bora-201	4.459597	2.0784216	0.0097249	muscle-pre.1
Knl1-201	4.337489	2.0168455	0.01410105	muscle-pre.1
Gm10184-201	4.077346	2.0365045	0.03458265	muscle-pre.1
Prc1-206	4.077346	1.9713358	0.03458265	muscle-pre.1
Nuf2-201	4.0189466	1.7332197	0.03933754	muscle-pre.1
Ska3-201	3.9977105	1.7515329	0.04120397	muscle-pre.1
Dbf4-203	3.9924014	1.7725917	0.04120397	muscle-pre.1
Ttk-201	3.9711652	1.7023541	0.04411713	muscle-pre.1
Nusap1-201	3.8543663	1.8275551	0.06481907	muscle-pre.1
Oip5-202	3.8331301	1.658908	0.06689453	muscle-pre.1
Cdca5-201	3.7694216	1.6803302	0.08074415	muscle-pre.1
Ndc80-201	3.7534943	1.5409368	0.08327797	muscle-pre.1
Ncapg-201	3.6632407	1.7917058	0.10843686	muscle-pre.1
Knstrn-206	3.6526225	1.5931767	0.1113843	muscle-pre.1
Entr1-201	3.6366954	2.4724135	0.11356238	muscle-pre.1
Coasy-201	3.6207683	1.5757657	0.11752356	muscle-pre.1
Kif11-201	3.578296	1.4158559	0.12799604	muscle-pre.1
Poc1a-201	3.5570598	1.6891567	0.1306336	muscle-pre.1
Cip2a-201	3.5570598	1.7929885	0.1306336	muscle-pre.1
Prc1-202	3.5411327	1.7184653	0.13558875	muscle-pre.1
Ckap2-201	3.4243336	1.612998	0.18065494	muscle-pre.1
Kpna2-	3.4243336	2.1289346	0.18065494	muscle-pre.1
Terf1-203	3.3977883	1.4129584	0.18975081	muscle-pre.1
Cenph-201	3.3871703	1.2960953	0.19540598	muscle-pre.1
Nek2-201	3.3712432	2.112818	0.20514444	muscle-pre.1
Ube2t-201	3.3128438	1.4995973	0.23471287	muscle-pre.1
Cdca2-204	3.3022256	2.015715	0.23471287	muscle-pre.1
Tpx2-202	3.3022256	2.217055	0.23471287	muscle-pre.1
Cenpl-203	3.2916076	1.6752703	0.2359557	muscle-pre.1
Psrc1-201	3.2916076	1.8760694	0.2359557	muscle-pre.1
Daxx-201	3.2703714	1.6053554	0.25040066	muscle-pre.1
Mbip-201	3.2013538	1.4214157	0.29999152	muscle-pre.1
Gemin6-	3.1854267	1.5211272	0.31011324	muscle-pre.1
Bub1b-201	3.1854267	1.5196954	0.31011324	muscle-pre.1
Ptma-	2.5917933	1.2528777	0.98456262	muscle-pre.2
Dcp2-201	2.62302	1.5583023	0.98456262	muscle-pre.2
Tmem30b-201	2.629265	2.4316359	0.98456262	muscle-pre.2
Slc25a5-	2.6729822	1.4240137	0.98456262	muscle-pre.2
Tpm1-213	2.6854727	2.0807443	0.98456262	muscle-pre.2
Ldhb-	2.5730577	1.3159151	0.98456262	muscle-pre.2
Ndufaf5-201	2.5480764	1.3475776	0.98456262	muscle-pre.2
B9d2-203	2.4543972	1.2891717	0.98456262	muscle-pre.2
Nicn1-201	2.4606423	1.1927199	0.98456262	muscle-pre.2
Ddt-201	2.5106046	1.6233656	0.98456262	muscle-pre.2
Rrm2-	2.5230954	1.7915854	0.98456262	muscle-pre.2
Srsf7-205	3.085171	2.4485793	0.98456262	muscle-pre.2
Hnrnpc-	3.2350578	2.2577627	0.98456262	muscle-pre.2
Tipin-205	3.322492	1.6066203	0.98456262	muscle-pre.2
Mpg-201	2.997737	2.0498493	0.98456262	muscle-pre.2
Riox1-201	2.7042086	1.5094658	0.98456262	muscle-pre.2
Trappc1-202	2.822869	1.5575259	0.98456262	muscle-pre.2
Rpl15-203	2.8353596	1.4044408	0.98456262	muscle-pre.2
Mcm5-205	2.9602652	1.7092898	0.98456262	muscle-pre.2
Cnih2-201	2.4356613	1.725709	0.98456262	muscle-pre.2

Slc38a4-201	5.233487	2.4577858	0.00054723	myoblast
Klhl13-	4.7577157	2.22412	0.00483094	myoblast
Spry1-	4.5198298	2.4397829	0.01077894	myoblast
Fxyd6-201	4.219887	1.7706419	0.03446158	myoblast
Spry1-201	3.992344	2.3537247	0.07450213	myoblast
Myf5-201	3.8734012	2.027822	0.10957714	myoblast
Cited1-202	3.837201	2.4036095	0.1228185	myoblast
Shisal2a-201	3.7130868	3.7927008	0.1638416	myoblast
Arhgap29-	3.6872296	2.3747003	0.17211266	myoblast
Fgfr4-201	3.6562011	3.2005439	0.18045416	myoblast
Gart-202	3.6044867	1.6802435	0.20578904	myoblast
Entpd4-206	3.5372581	1.9686908	0.25469451	myoblast
Prkci-201	3.5217438	1.4278562	0.25905077	myoblast
Nfib-202	3.464858	1.5686892	0.2855887	myoblast
Dlk1-208	3.4700296	1.7494532	0.2855887	myoblast
Arfip2-206	3.464858	1.4423202	0.2855887	myoblast
Pitx3-201	3.4700296	1.4916306	0.2855887	myoblast
Tmem246-201	3.4338295	2.5512786	0.31464756	myoblast
P3h2-201	3.4183152	3.4937754	0.32729867	myoblast
Top2b-201	3.402801	1.6115663	0.3291366	myoblast
Fzd4-201	3.3924582	3.6997597	0.33621353	myoblast
Mta1-204	3.3355725	1.7063799	0.37617183	myoblast
Entpd4b-204	3.2786865	2.0241003	0.43490972	myoblast
Cnot10-201	3.2062867	1.5141709	0.509456	myoblast
Ddx51-201	3.139058	2.1518767	0.57388791	myoblast
Pdgfa-202	3.1338866	1.3962227	0.57388791	myoblast
Gm266-201	3.1235437	3.5360732	0.58784153	myoblast
Rps8-203	2.9632294	1.5586787	0.81203166	myoblast
Zfp580-201	2.9270294	1.3217518	0.84436227	myoblast
Parm1-201	2.8960009	1.7470729	0.90978745	myoblast
Tnnt1-203	5.629205	3.4453943	4.47E-05	myocyte
Bin1-206	5.5319405	3.338935	5.21E-05	myocyte
Slc25a3-206	5.3921223	5.3179917	7.36E-05	myocyte
Eno1-	5.40428	5.4388027	7.36E-05	myocyte
Tnnt1-	5.300936	2.7793067	0.0001066	myocyte
Tmed2-	5.167197	7.7333364	0.00018515	myocyte
Lsm6-204	5.1064067	3.917087	0.00022094	myocyte
Cdkn1c-	5.045616	3.342858	0.00028478	myocyte
Hacd1-203	5.0213	5.6134887	0.00031012	myocyte
Serinc2-203	4.7902956	4.1015873	0.0008218	myocyte
Tnnt1-206	4.6626353	2.8404799	0.00133959	myocyte
Vasp-201	4.644398	3.4756563	0.00140255	myocyte
Cdkn1c-	4.620082	3.1417873	0.0015143	myocyte
Tnnt1-201	4.614003	3.05506	0.00153878	myocyte
Cdkn1c-	4.510659	2.4062722	0.0024221	myocyte
Shisa2-201	4.3586826	2.6834664	0.00435296	myocyte
Vma21-203	4.3282876	6.450808	0.00483619	myocyte
Pnmal2-201	4.3222084	3.6279802	0.00491796	myocyte
Acta2-	4.273576	3.1073308	0.00587162	myocyte
Ube2d3-	4.267497	5.277565	0.00597231	myocyte
Lsm6-	4.2310224	4.984635	0.00681955	myocyte
Vdac3-	4.2127852	3.1398304	0.00725082	myocyte
Lgmn-	3.949313	5.6585793	0.03270791	macrophage
Mpp1-201	3.9312797	4.1905756	0.03270791	macrophage
Lcp2-201	3.859146	5.2996893	0.03270791	macrophage
Cebpd-201	3.859146	3.9262567	0.03270791	macrophage
Atp6v0a1-	3.859146	4.9282146	0.03270791	macrophage
Zfp36-202	3.859146	4.560254	0.03270791	macrophage
Mt1-201	3.859146	4.929718	0.03270791	macrophage

Cd53-201	3.8771794	5.9794636	0.03270791	macrophage
Fam49b-207	3.895213	4.1351047	0.03270791	macrophage
Zfp36-201	3.895213	5.6155725	0.03270791	macrophage
Psmb10-201	3.9132462	4.556701	0.03270791	macrophage
Chd9-	4.003413	6.0899177	0.03270791	macrophage
Dab2-203	4.057513	8.815356	0.03270791	macrophage
Lgmn-202	4.057513	6.673731	0.03270791	macrophage
Fyb-210	4.057513	29.61367	0.03270791	macrophage
Ifngr1-201	4.0394797	4.342029	0.03270791	macrophage
Sesn1-201	4.003413	4.205529	0.03270791	macrophage
Grn-201	4.003413	5.042139	0.03270791	macrophage
Dab2-202	3.859146	4.1632442	0.03270791	macrophage
Cndp2-202	4.0214467	4.1234236	0.03270791	macrophage

Table 4.10. Single-cell batch and sequencing platform information

cell_ID	celltype	batch	Sequel	HudsonAlpha Sequel II	UCI Sequel II
17329_C5	muscle-pre	pool1	X	X	
17333_G6	muscle-pre	pool1	X	X	
18042_A7	muscle-pre	pool1	X	X	
18046_E3	muscle-pre	pool1	X	X	
18046_E8	muscle-pre	pool1	X	X	
18088_G2	muscle-pre	pool1	X	X	
18088_G8	EMP	pool1	X	X	
18252_B3	macrophage	pool1	X	X	
18254_D10	EMP	pool1	X	X	
18255_E8	myocyte	pool1	X	X	
18256_F2	myoblast	pool1	X	X	
18264_B9	muscle-pre	pool1	X	X	
18267_E11	muscle-pre	pool1	X	X	
18311_A9	myoblast	pool1	X	X	
18312_B8	myoblast	pool1	X	X	
18312_B12	macrophage	pool1	X	X	
18314_D6	myocyte	pool1	X	X	
19907_B3	muscle-pre	pool1	X	X	
19916_C5	EMP	pool1	X	X	
19917_D1	myocyte	pool1	X	X	
20026_A12	muscle-pre	pool1	X	X	
20032_G2	EMP	pool1	X	X	
20034_B3	EMP	pool1	X	X	
20034_B6	myocyte	pool1	X	X	
20038_F1	myocyte	pool1	X	X	
20040_B1	myocyte	pool1	X	X	
20040_B6	myocyte	pool1	X	X	
20041_C1	myoblast	pool1	X	X	
20042_D2	myocyte	pool1	X	X	
20043_E2	myoblast	pool1	X	X	
20046_C4	macrophage	pool1	X	X	
20047_D6	myocyte	pool1	X	X	
18044_C5	muscle-pre	pool2			X
18044_C6	muscle-pre	pool2			X
18044_C9	muscle-pre	pool2			X
18048_G2	muscle-pre	pool2			X
18048_G4	muscle-pre	pool2			X
18048_G8	muscle-pre	pool2			X
17328_B5	muscle-pre	pool2			X
17328_B8	muscle-pre	pool2			X
17329_C8	muscle-pre	pool2			X
18251_A3	myoblast	pool2			X
18251_A5	myoblast	pool2			X
18252_B8	myocyte	pool2			X
18252_B11	myoblast	pool2			X
18252_B12	myocyte	pool2			X

cell_ID	celltype	batch	Sequel	HudsonAlpha Sequel II	UCI Sequel II
18254_D2	myocyte	pool2			X
18254_D12	myoblast	pool2			X
18255_E2	myoblast	pool2			X
18255_E3	macrophage	pool2			X
18255_E4	myoblast	pool2			X
18255_E5	myoblast	pool2			X
18255_E11	myoblast	pool2			X
18255_E12	macrophage	pool2			X
18257_G6	myoblast	pool2			X
18258_A7	muscle-pre	pool2			X
18263_A2	myocyte	pool2			X
18265_C6	muscle-pre	pool2			X
18270_A4	myoblast	pool2			X
18270_A11	myoblast	pool2			X
18271_B2	myoblast	pool2			X
18274_E3	myoblast	pool2			X
18312_B2	macrophage	pool2			X
18313_C10	myocyte	pool2			X
18316_F6	myocyte	pool2			X
18317_G2	myoblast	pool2			X
18317_G10	myocyte	pool2			X
20026_A5	myoblast	pool2			X
20026_A6	myoblast	pool2			X
20026_A9	muscle-pre	pool2			X
20031_F8	muscle-pre	pool2			X
20032_G5	muscle-pre	pool2			X
20039_A1	myoblast	pool2			X
20042_D4	myocyte	pool2			X
20044_A10	myocyte	pool2			X
20044_A12	myoblast	pool2			X
20048_E11	myoblast	pool2			X
20048_E12	myoblast	pool2			X
19914_A12	muscle-pre	pool2			X
20043_E4	macrophage	both	X	X	X
18316_F2	macrophage	both	X	X	X

4.6 References

1. Uzman, A., Lodish, H., Berk, A., Zipursky, L. & Baltimore, D. Molecular Cell Biology (4th edition) New York, NY, 2000, ISBN 0-7167-3136-3. *Biochem. Mol. Biol. Educ.* (2000). doi:10.1016/S1470-8175(01)00023-6
2. Davidson, E. H. The regulatory genome: gene regulatory networks in development and evolution. *Developmental Biology* (2006). doi:10.1016/j.ydbio.2007.08.009
3. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
4. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (80-.).* (2008). doi:10.1126/science.1158441
5. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* (2017). doi:10.1038/ncomms14049
6. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* (2015). doi:10.1016/j.cell.2015.05.002
7. Guo, G. *et al.* Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Dev. Cell* (2010). doi:10.1016/j.devcel.2010.02.012
8. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* (2014). doi:10.1038/nbt.2859
9. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* (2008). doi:10.1038/nature07509
10. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* (2010). doi:10.1038/nbt.1621
11. Early, P. *et al.* Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell* (1980). doi:10.1016/0092-8674(80)90617-0
12. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.* (2003). doi:10.1146/annurev.biochem.72.121801.161720
13. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: Towards a cellular code. *Nature Reviews Molecular Cell Biology* (2005). doi:10.1038/nrm1645
14. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology* (2017). doi:10.1038/nrm.2017.27
15. Buckingham, M. *et al.* The formation of skeletal muscle: From somite to limb. *Journal of Anatomy* (2003). doi:10.1046/j.1469-7580.2003.00139.x
16. Martin, P. Tissue patterning in the developing mouse limb. *Int. J. Dev. Biol.* (1990). doi:10.1387/ijdb.1702679
17. Christ, B. & Brand-Saberi, B. Limb muscle development. *Int. J. Dev. Biol.* (2002). doi:10.1387/ijdb.12455628
18. Lagha, M. *et al.* Transcriptome analyses based on genetic screens for Pax3 myogenic targets in the mouse embryo. *BMC Genomics* (2010). doi:10.1186/1471-2164-11-696
19. Biressi, S. *et al.* Intrinsic phenotypic diversity of embryonic and fetal myoblasts is

revealed by genome-wide gene expression analysis on purified cells. *Dev. Biol.* (2007). doi:10.1016/j.ydbio.2007.01.016
20. Chal, J. & Pourquié, O. Making muscle: Skeletal myogenesis in vivo and in vitro. *Development (Cambridge)* (2017). doi:10.1242/dev.151035
21. Hasty, P. *et al.* Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. *Nature* (1993). doi:10.1038/364501a0
22. Venuti, J. M., Morris, J. H., Vivian, J. L., Olson, E. N. & Klein, W. H. Myogenin is required for late but not early aspects of myogenesis during mouse development. *J. Cell Biol.* (1995). doi:10.1083/jcb.128.4.563
23. He*, P. Williams, B. A.*, Trout, D., Marinov, G. K., Amrhein, H., Berghella, L., Goh, S., Plajzer-Frick, I., Afzal, V., Pennacchio, L. A., Dickel, D. E. , Visel, A., Ren, B. , Hardison, R. C., Zhang, Y. and W. B. J. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Accept. Publ. Nat.* (2020).
24. Chapman, A. B., Knight, D. M., Dieckmann, B. S. & Ringold, G. M. Analysis of gene expression during differentiation of adipogenic cells in culture and hormonal control of the developmental program. *J. Biol. Chem.* (1984).
25. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* (2018). doi:10.1186/s13059-017-1382-0
26. Miller, J. B. & Stockdale, F. E. Developmental origins of skeletal muscle fibers: Clonal analysis of myogenic cell lineages based on expression of fast and slow myosin heavy chains. *Proc. Natl. Acad. Sci. U. S. A.* (1986). doi:10.1073/pnas.83.11.3860
27. Harris, A. J., Duxson, M. J., Fitzsimons, R. B. & Rieger, F. Myonuclear birthdates distinguish the origins of primary and secondary myotubes in embryonic mammalian skeletal muscles. *Development* (1989).
28. Biressi, S., Molinaro, M. & Cossu, G. Cellular heterogeneity during vertebrate skeletal muscle development. *Developmental Biology* (2007). doi:10.1016/j.ydbio.2007.06.006
29. Hutcheson, D. A., Zhao, J., Merrell, A., Haldar, M. & Kardon, G. Embryonic and fetal limb myogenic cells are derived from developmentally distinct progenitors and have different requirements for β-catenin. *Genes Dev.* (2009). doi:10.1101/gad.1769009
30. Yang, J. *et al.* RBM24 Is a major regulator of muscle-specific alternative splicing. *Dev. Cell* (2014). doi:10.1016/j.devcel.2014.08.025
31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
32. Wyman, D. & Mortazavi, A. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics* **35**, 340–342 (2019).
33. Wyman, D. *et al.* A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv* (2019). doi:10.1101/672931
34. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
35. Anvar, S. Y. *et al.* Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, 46 (2018).

CHAPTER 5

Future Directions

Chapter 5

Future Directions

In the previous chapters, I outlined how the novel long-read analysis methods I developed may be used to characterize isoform-level mRNA expression in both bulk RNA samples and in single cells. I first described TranscriptClean, a tool for correcting mismatches, small indels, and noncanonical splice junctions in long-read alignments using the reference genome. Next, I introduced TALON, a novel, technology-agnostic pipeline for annotation and quantification of long-read transcriptomes. Lastly, I demonstrated how TranscriptClean and TALON can be applied to deeply sequenced long-read single-cell data to better understand the role of isoform switching in the developing embryonic mouse limb bud.

I anticipate many ways to build on this work. To start, it would be interesting to move beyond human cell lines and examine isoform-level differences in human tissues using long-read sequencing. The ENTEx collaboration between the GTEx[1] and ENCODE[2] consortia presents an exciting opportunity to do just that. This project entails deep profiling of tissues from four individual human donors by a variety of genomics assays, including but not limited to ATAC-seq, ChIP-seq, eCLIP, short-read RNA-seq, and PacBio long-read sequencing[3]. In addition, work is underway to assemble phased, diploid personal genomes for the four ENTEx donors. So far, our laboratory has performed PacBio long-read sequencing on the transcriptomes of four donor tissue samples. From these data, it is possible to quantify isoform differences between tissues and to examine how isoform

expression varies in the same tissue across the ENTEx individuals. Ultimately, an important goal would be to understand the effects of personal genetic variation on isoform-level expression. Given the small sample size available in this study, it may be difficult to draw broad conclusions in this regard. Nevertheless, it is possible to envision integrative analyses on the ENCODE data at large that would ask variations of this question. For instance, eCLIP, RNA-seq, and PacBio data from the same cell lines could be combined to examine the effect of particular RNA binding proteins on alternative splicing. Personal variation within RNA binding protein motifs could play an important role here. In addition, the availability of phased genomes could allow for study of allele-specific isoform expression, which I will discuss further in the next section.

Because current long-read platforms produce data with single-molecule resolution, it is in principle possible to tell whether each read comes from the maternal or the paternal copy of the gene. This introduces the exciting possibility of quantifying allele-specific expression on the isoform level, and perhaps to better understand the relationship between genetic variation and splicing patterns. Tilgner et al. pioneered this type of analysis in 2014, developing a principal component analysis-based approach to assigning individual PacBio circular consensus reads to the maternal or paternal haplotype of GM12878[4]. At the time, the high long-read error rate and low throughput of the technology meant that calling differential allelic expression was still a challenge. Improvements in the technology since may make this more feasible today. In practice, such analyses will of course be difficult to conduct for samples lacking a phased personal genome. However, it may be feasible to modify the TALON pipeline to record phased isoform abundance

estimates for well-studied cell lines such as GM12878 or for the previously mentioned ENTEx personal genomes.

The successful application of long-read sequencing to single cells has opened the door to a variety of interesting possibilities. Since alternative splicing is known to be an important factor in many diseases including Alzheimer's, a natural next step is to characterize single-cell transcriptomes from diseased and healthy tissues and look for isoform-level differences with pathogenic implications. Combining such an analysis with matching personal genome data from the patients could also help illuminate the relationship between genetic variation and alternative splicing in the disease context. As shown in Chapter 4, the TALON pipeline can readily be applied to single-cell datasets.

Although long-read technologies offer many advantages over short-read sequencing, they still pose significant challenges of their own. It is important to develop both computational and experimental solutions to these issues so that we can more accurately quantify transcriptomes on the isoform level. A major challenge for cDNA protocols is to fully sequence long transcripts, such as those measuring over 5 kb in length. More processive reverse transcriptases are needed in order to make this possible. In addition, internal priming is a widespread experimental artifact that can lead to truncated transcripts in protocols that use poly-(A) selection. In Chapter 3, I outlined a computational approach used by TALON to identify and flag reads believed to come from internal priming. However, it may also be possible to reduce these artifacts on the experimental side by using a reverse-transcriptase that works at elevated temperatures (i.e. 50°C or higher). The

goal here would be to force more stringent hybridization conditions and reduce off-target priming events. Computational measures may also be taken to help identify full-length transcripts provided that orthogonal data types are available. For instance, it would be possible to adapt the CAGE and poly-(A) motif analysis outlined in Chapter 3 to use 5' and 3' end support as a necessary condition when filtering novel transcript models in TALON. This could help remove artifactual suffix ISMs resulting from RNA degradation or incomplete reverse-transcription. Finally, there is the matter of artifactual long-read splice junctions. TranscriptClean was designed at the outset to correct exclusively noncanonical splice junctions, since these are highly likely to be the result of sequencing errors rather than biological novelty. However, our SIRV analysis in Chapter 3 indicates that not all artifactual splice junctions are noncanonical – we see unexpected SIRV junctions with canonical motifs as well. In light of this, it may be useful in the future to consider the level of splice junction support during transcript filtering as well, making use of short-read and annotation support information where available. Overall, quality control measures such as those suggested here can further hone the accuracy of third-generation sequencing platforms and improve their utility for isoform-level analyses on the bulk and the single-cell level.

References

1. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).
2. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
3. Stranger, B. E. *et al.* Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nature Genetics* (2017). doi:10.1038/ng.3969
4. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci.* (2014). doi:10.1073/pnas.1400447111