Characterizing isoform diversity in endothelial cells via proteogenomic methods that integrate long-read RNA-sequencing and mass-spectrometry based proteomics

Madison Melody Mehlferber
Silver Spring, Maryland

Bachelor of Science Bioinformatics, Saint Vincent College, Latrobe, Pennsylvania, 2019
Master of Science in Biological and Physical Sciences, University of Virginia, Charlottesville, Virginia, May 2023

A Dissertation Presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biochemistry & Molecular Genetics
University of Virginia
November 1st, 2024

Abstract

The process of alternative splicing enables one gene to produce multiple transcript isoforms that can produce proteins with distinct functions, expanding the diversity of the human proteome. The individual isoforms resulting from alternative splicing provide discrete levels of regulation that can modulate distinct cell phenotypes or cell fates. The evolution of RNA-sequencing technology has contributed to illuminating the prevalence of alternative splicing events, revealing that 95% of human genes are subjected to this process. However, the availability of comprehensive methods that delineate both transcript isoforms and associated protein isoforms remain limited. As a result, the isoform atlas remains incomplete for many tissue types, hindering knowledge of the degree at which alternative splicing impacts the proteome. In my dissertation work, I utilized innovative approaches to more effectively profile isoforms and their associated protein isoforms. My research benefitted from advancements in long-read RNA-sequencing technology, increasing resolution of the human transcriptome. I highlight in my first chapter, the development of an approach capitalizing on such RNA-sequencing advancements and integrating it with proteomics to create an isoform atlas. In my second chapter, I describe an application of this method to characterize the isoform landscape within endothelial cells, discovering novel protein isoforms for key markers of endothelial cell identity. I then review the roles of isoforms in directing cell fate decisions and their influence within gene regulatory networks. In my next chapter, I describe the knowledge of the potential roles of isoforms in development and how increases in resolution and throughput of long-read RNA-sequencing technologies were enabled by the introduction of a novel concatenation technique called MAS-Iso-Seq. These advances enabled the construction of an isoform atlas describing transcriptome dynamics during the process of early endothelial cell differentiation. Finally, I offer my perspectives on the state of the transcriptome analysis field and describe potential future directions for how such analysis can support precision medicine applications. Overall, my dissertation work supports enhanced profiling of isoform populations with specific focus on construction of an isoform atlas within endothelial cell types.

Signatures

_____

Gloria M. Sheynkman, PhD

_____

Stefan Berkiranov, PhD

_____

Karen K. Hirschi, PhD

_____

Mete Civelek, PhD

_____

Muge Kuyumcu-Martinez, PhD

_____

John Chance Luckey, PhD, MD

Dedications

To my Mom and Dad. I love you more than words can ever express. Thank you for being my biggest advocates and supporters. Thank you for always believing in me and empowering me to chase my dreams. I am so proud to be your daughter, you will forever be my inspiration. You've provided me with opportunities that some could only dream of, I'm beyond grateful to have had the privilege to live them. You've championed me through it all. I hope you always know how much its meant to have you all in my corner, because it's something pretty extraordinary, and in fact it's the greatest privilege of a lifetime. Thank you for all that you do and being extraordinary. I'm everything because of you.

To Gloria, thank you for entrusting me to be your lab's first graduate student at the University of Virginia. Thank you for providing the unique experience that comes with this responsibility allowing me to contribute to shaping and defining the early questions of the lab. I'm forever thankful for the opportunities you've provided me and the contributions you have made to my scientific development. Thank you for instilling the curiosity and drive to accomplish and tackle the difficult questions.

To Erin, thank you for always being my advocate and supporter. I knew from the very moment I joined the lab that this was going to be something special and my goodness I was right. Thank you for not only teaching me all the New York sayings contributing to my vocabulary but sharing your knowledge and wisdom. I'm beyond honored to have learned mass-spectrometry from you and feel privileged to have joined the Hunt Lab group messages. Thank you for sharing all the moments over the years the highs, the lows and everything in between along the way. Thank you for taking care of me and stepping in to be lab mom when needed, you wear it incredibly well. Your compassion and support are something gigantic, shining bright even in the smallest gestures.

To current Sheynkman Lab members, Jennifer, Micah, Bond, Vasilii, Erin, Leon and Emily thank you for being a part of my Ph.D. journey. I look back fondly at the memories from conference trips we've shared and the daily interactions in the write up space that was a part of my journey. I'm so proud to see the scientists that you are becoming, and I hope you all know how bright your future is. Thank you for all the laughs, the support and the pleasure of growing as a scientist with you all. It has been so fun to see the lab grow over the years, and I am so honored to stand among what I know will be a legacy of amazing scientists.

To Karen, thank you for making my graduate career so bright. Your lab offered one of my first homes and my first friendships at UVA and I am so appreciative. Thank you for welcoming me and sharing your wisdom and knowledge. Thank you for introducing me to vascular biology and providing me an opportunity to flourish and embracing our interdisciplinary approaches.

Thank you for your support and your commitment to my training and development as a scientist and my future career.

To the rest of my committee members, Mete, Stefan, John, Muge and David, thank you for your insightful comments and support through my training. Your contribution and support have made me a better scientist. Your commitment to my training and development will is something I will appreciate far beyond my graduate career. Thank you for your valuable contributions to my training and thank you for allowing me the privilege of your wisdom and training.

To my friends, thank you for your unwavering support, the laughs and all the joy. Thank you for making this journey so much fun. The memories made with you all will last a lifetime and will always make me smile.

To Charlottesville, thank you for being the vibrant town that I have been able to call home for the last five years. Choosing this campus and this university remains one of the best decisions I have ever made. I will forever cherish the instant charm of this beautiful town that lured me in during my interviews as a young scientist. Cheers to the beautiful memories.

To my previous mentors, Dr. Michael Sierk from Saint Vincent College, Dr. Ken Yamada and Dr. Shaohe Wang from NIDCR and Dr. Matt Kelley from NIDCD thank you for providing my scientific foundation. Thank you for teaching me what good science is both at the bench and beyond.

Attributions

The work presented in this thesis is the result of collaborative efforts between myself and the very many talented individuals and scientists that I have had the pleasure to work with who have supported my scientific development. Much of the text found in this thesis is the product of text adapted from previous work or publications. Throughout this document, I will highlight the contributions of the scientists that were involved in the work presented in this dissertation and the associated publications described.

Chapter 2:

This chapter is adapted from:

Miller RM, Jordan BT, **Mehlferber MM**, Jeffery ED, Chatzipantsiou C, Kaur S, Millikin RJ, Dai Y, Tiberi S, Castaldi PJ, Shortreed MR, Luckey CJ, Conesa A, Smith LM, Deslattes Mays A, Sheynkman GM. Enhanced protein isoform characterization through long-read proteogenomics. Genome Biol. 2022 Mar 3;23(1):69. doi: 10.1186/s13059-022-02624-y. PMID: 35241129; PMCID: PMC8892804.

This study and concept were created by a very talented senior graduate student I had the pleasure to work with early in my career, Rachel Miller in collaboration with Gloria M. Sheynkman. I had the terrific opportunity to be a part of the codeathon weekend led by Dr. Rachel Miller and Dr. Anne-Deslattes Mayes that provided the foundation behind the development of the long-read proteogenomics pipeline. My collaboration to this process and the educational experience this offered early in my graduate career was hugely impactful in solidifying my desire to join the Sheynkman Lab and continue efforts related to characterizing isoform landscapes. Overall, GMS designed the study and supervised the project in addition to LMS, ADM, AC, MRS, and PC. RMM, BTJ, CC, SK, RJM, MRS, ADM, and GMS developed the open-source computational pipeline. Specifically highlighted in this paper was work greatly supported by the mentorship from Erin D. Jeffery teaching me about all aspects of mass-spectrometry and supervising me in learning the process of novel peptide discovery principles. MMM and EDJ performed the novel peptide analysis. BTJ, CC, YD, and ADM contributed to the analysis reproducibility, data curation, and design of the workflow. CJL performed the biological analysis of Jurkat isoforms. BTJ, GMS, and AC worked on the SQANTI Protein classification scheme.

Chapter 3:

This chapter is based on work described in the following publication:

Mehlferber MM, Jeffery ED, Saquing J, Jordan BT, Sheynkman L, Murali M, Genet G, Acharya BR, Hirschi KK, Sheynkman GM. Characterization of protein isoform diversity in human umbilical vein endothelial cells via long-read proteogenomics. RNA Biol. 2022 Jan;19(1):1228-1243. doi: 10.1080/15476286.2022.2141938. PMID: 36457147; PMCID: PMC9721438.

GMS and MMM conceived of the project. GMS designed the study and supervised the project, along with KKH. MMM was involved in data collection, data analysis, interpretation, and conclusions, with discussions with GMS. LS performed the long-read RNA-seq experimental analysis. BTJ performed computational analysis, including long-read RNA-seq and proteomics analysis. MMM and EDJ performed the novel peptide analysis. JS and MM performed computational analysis of novel isoforms. BTJ, MMM and GMS contributed to analysis reproducibility, data curation and design of the workflow for the data described in this paper. MMM and GMS wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Chapter 4:

This chapter was adapted from:

Mehlferber, M.M., Kuyumcu-Martinez, M., Miller, C.L. et al. Transcription Factors and Splice Factors—Interconnected Regulators of Stem Cell Differentiation. Curr Stem Cell Rep 9, 31–41 (2023). https://doi.org/10.1007/s40778-023-00227-2

MMM and GMS conceived of the project and the scope of the review. MMM, CLM provided additional support and examples to supplement the content. MMM designed the figure using Biorender.com

Chapter 5:

The content of this chapter is adapted from:

Transcript detection and quantification using Kinnex Full-length RNA Sequencing Data - David Wissel, University of Zurich, Joint work with **Madison M. Mehlferber** with joint supervision by Gloria Sheynkman (UVA) and Mark D. Robinson (University of Zurich) https://programs.pacb.com/l/1652/2024-03-08/44gtlx

And

Kinnex full-length RNA kit for isoform sequencing - https://www.pacb.com/wp-content/uploads/Application-note-Kinnex-full-length-RNA-kit-for-isoform-sequencing.pdf

This chapter contains unpublished data in collaboration with David Wissel and Mark D. Robinson from the University of Zurich in Switzerland. Madison M. Mehlferber wrote and researched the background on PacBio long-read RNA-sequencing describing the advancements in technology supporting the goals of isoform discovery. MMM also conceived of the time series RNA experiment for isoform detection within primordial EC cells, performing the experimental work with Vasilii Pavelko to generate the samples used for analysis both in this chapter and within Chapter 6. DW and MDR provided expertise on the proper software and analysis needed to effectively analyze and support time-series long-read RNA sequencing data. Figures are adapted and noted accordingly where they were utilized in David Wissel's webinar analyzing approaches for transcript discovery and quantification, with the new Kinnex platform as highlighted in PacBio spotlights on applications series. The dataset described here and further

described in Chapter 6 is the product of this collaboration with this dataset serving to support multiple analysis goals.

Chapter 6

This chapter contains preliminary unpublished data:

Madison M. Mehlferber, David Wissel, Vasilii Pavelko, Elizabeth Nelson, Emily Watts-Whitehead, Erin D. Jeffery Mark D. Robinson, Leon Sheynkman, Gloria M. Sheynkman

This chapter contains preliminary and unpublished data in collaboration with David Wissel and Mark D. Robinson at the University of Zurich. MMM conceived of the time series study with experimental guidance from EN and assistance with sample collection provided by VP. MMM conceived of the analysis and received support and supervision by DW and MDR. DW and MDR advised and provided expertise on the optimal approaches for handling time-series data analysis and analyzing differential transcript expression profiles. MMM also sought guidance from Karen K. Hirschi in describing the most impactful examples for the vascular biology field found within the dataset.

Chapter 7

This chapter provides a summary of the chapters enclosed within this thesis. In addition, I offer my perspectives on the transcriptome analysis field and potential future directions for the field describing mentions of new technologies and approaches to support successful biomedical breakthroughs.

Table of Contents:

List of Figures:

List of Tables:

Abbreviations:

AA	Amino acid
AS	Alternative splicing
bp	Base pair
CCS	Circular consensus reads (generated via PacBio technologies)
cDNA	Complimentary deoxyribonucleic acid
CPM	Counts per million
DDA	Data-dependent acquisition
DGE	Differential gene expression
DNA	Deoxyribonucleic acid
DTE	Differential transcript expression
DTU	Differential transcript usage
EC	Endothelial cells
ENCODE	Encyclopedia of DNA elements
FDR	False discovery rate
FSM	Full-splice match
GO	Gene ontology
HiFi	High-fidelity reads resulting from PacBio sequencing
HPLC	High-performance liquid chromatography
HUVECs	Human Umbilical Vein Endothelial Cells
iPSCs	Induced Human Pluripotent Stem Cells
kbp	Kilobase pairs (metric for measuring transcripts from RNA-sequencing)
LC-MS	Liquid chromatography- mass spectrometry
LR-seq	Long-read RNA sequencing
M/z	Mass to charge ratio

mRNA	Messenger ribonucleic acid
MS	Mass Spectrometry
NIC	Novel in catalog
NNC	Novel not in catalog
ONT	Oxford Nanopore Technologies
ORFs	Open-reading frames
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
pFSM	Protein full-splice match
pNIC	Protein novel in catalog
pNNC	Protein novel not in catalog
PSI	Percent spliced in
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
SF	Splice factor
SMRT-sequencing	Single-molecule real time sequencing
TF	Transcription factor
ZMW	Zero-mode waveguide

Chapter 1 Introduction

1.1 Background on isoforms

1.1.1 Defining isoforms in the human genome

The central dogma underlies the foundation of biology in which genetic information exists in a unidirectional continuum where DNA stores genetic information that encode genes. From this, genetic information can be transferred via transcription, to produce RNA molecules. These RNA molecules are then translated into proteins that form the functional and active units of the cell which perform their associated function as destined by their underlying genetic code **(Figure 1.1).** The human genome is comprised of about 20,000 protein-coding genes.

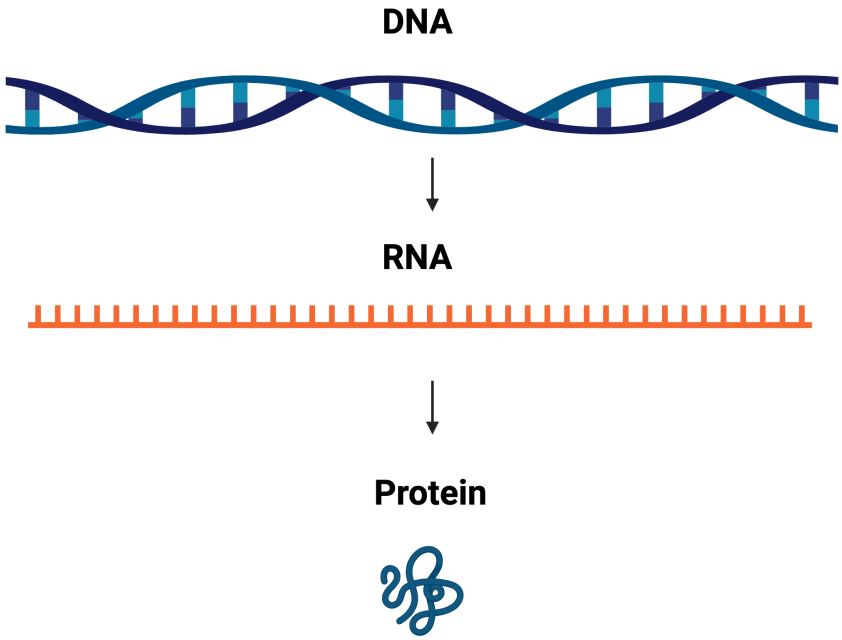


Figure 1.1 The central dogma of biology describes the flow of genetic information in living systems
Genetic information flows through a biological system and is stored in DNA. The process of transcription allows the genetic information to be copied within the nucleus of the cell into messenger RNA (mRNA) molecules. Once RNA is transcribed, it exits the nucleus and binds to ribosomes that translate the RNA code into proteins, the functional units of the cell.

However, an added complexity exists in this simplistic portrayal of the transfer of genetic information, revising the idea that one gene translates to one protein. Genes can give rise to multiple RNA products through the process of alternative splicing (AS), discovered by Philip Sharp and colleagues in the 1970s (Berget, Moore, and Sharp 1977). In this study, they focused on modeling the process of transcription where it was observed, the messenger ribonucleic acid (mRNA) strands produced via transcription by the adenovirus when hybridized to the genome did not match the DNA template sequence to which it was derived. Rather, the mRNA represented novel combinations of sequences that seemed to be "spliced together" from dis-contiguous parts of the DNA coding regions.

It was later understood that the pre-mRNA transcripts produced by an organism via transcription could be subject to the removal of intronic (non-coding regions) and that the exons (coding) can be ligated together to form a mature mRNA transcript isoform. These exons could be differentially included during AS to create distinct transcript products unique from the original DNA sequence that they were derived. The process of splicing is quite complex, requiring tight regulation to permit selective inclusion or exclusion of exons within the final mRNA transcript. The splicing machinery (the spliceosome) is used to facilitate this process and is comprised of over 200 proteins (Cvitkovic and Jurica 2013). To achieve sensitive selection of the proper splice sites, a cascade of binding events between splice-factors (SFs) and the spliceosome occurs on the pre-mRNA, regulating the appropriate exons to include. These binding events cause recruitment of the spliceosome, which catalyzes the joining of the selected exons. Additional details on this process are outlined in **Chapter 4.3.1.**

The process of AS is responsible for contributing to a major source of genetic variation (F. Shen et al. 2023). Rather than the human genome having simply the repertoire of functions achieved by 20,000 genes (Aebersold et al. 2018), it is predicted that there are at least 200,000 alternative transcripts with more being discovered (Adam Frankish et al. 2021). Additionally, AS contributes a major source of gene regulation, enabling refinement of the exact combinations of exons necessary for a particular function. These distinct transcript isoforms can be translated into proteins that may go on to produce functionally distinct proteins, or proteoforms (L. M. Smith, Kelleher, and Consortium for Top Down, Proteomics 2013) **(Figure 1.2)**.

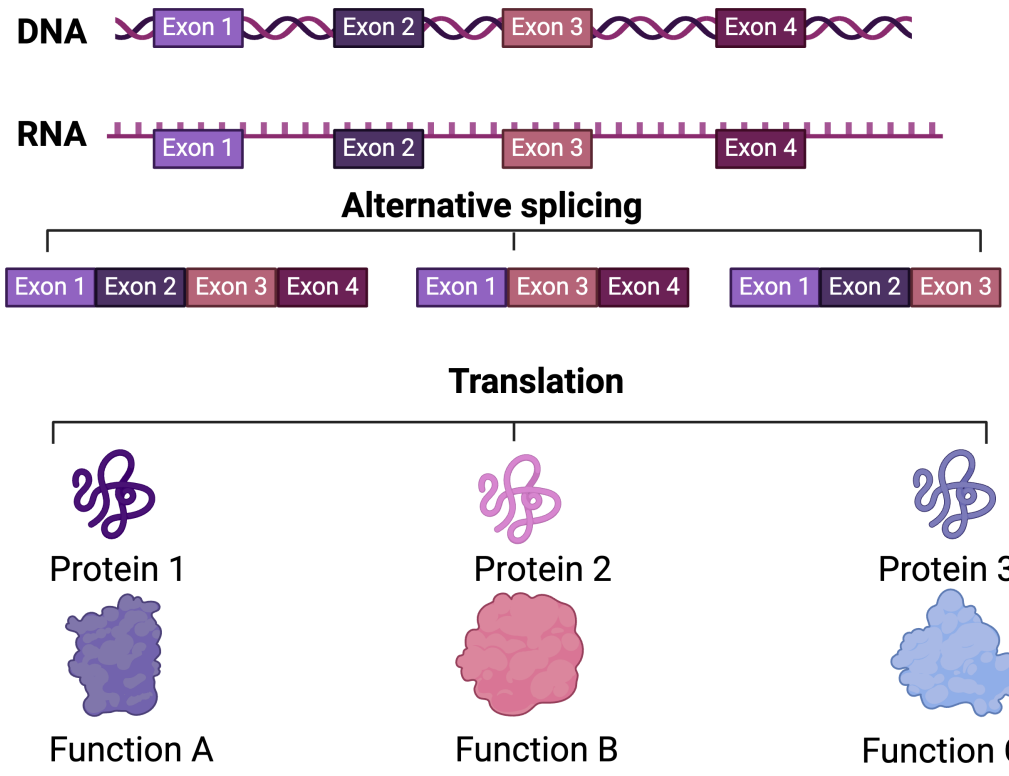


Figure 1.2 The process of alternative splicing contributes to proteomic diversity
DNA forms the template by which RNA is transcribed. During the process of transcription, the introns within the pre-mRNA are removed (spliced out) and the exons are ligated together into distinct arrangements forming a mature mRNA transcript. The mRNA generated from the alternative transcripts can be translated into a protein and can possess distinct functions depending on the unique combinations of sequence contained in the alternative protein product.

It is predicted that over 90% of the human genome undergoes AS, supporting diversification of the proteome (E. T. Wang et al. 2008; Yansheng Liu et al. 2017). While AS enables proteomic diversity, aberrant splicing is associated with 15% of human diseases that present based on heredity (Ibeh et al. 2024; R. Wang et al. 2023), and one-third of disease-causing mutations are attributed to aberrant RNA splicing patterns (Montes et al. 2019; Scotti and Swanson 2016).

AS can also act in a tissue-specific manner, generating splicing patterns yielding distinguishable transcriptional signatures to specific tissue types (Melé et al. 2015; E. T. Wang et al. 2008). The process of AS also is associated with defining developmental stages. Specific groups of isoforms might be expressed highly during early stages of development, with decreased expression of those isoforms associated with later developmental stages (Mazin et al. 2021). Important switches in these isoform expression ratios can govern cell differentiation processes and assimilation of tissue-specific properties (Fiszbein and Kornblihtt 2017). The

ability to detect and characterize AS-driven transcriptional signatures can help us understand the relationship between transcript isoforms and biological phenotypes.

1.2 Approaches for detection of RNA transcript isoforms

1.2.1 The origin of RNA-sequencing technologies

Our understanding of the prevalence of isoforms produced via AS has been greatly influenced by short-read RNA sequencing technology (Z. Wang, Gerstein, and Snyder 2009).

The evolution of RNA-sequencing has been defined by three major waves of technological advancements. The "first-generation sequencing" wave has been defined by the development of automated Sanger sequencing (**Figure 1.3**, "First-generation sequencing") (Mardis 2013). This technology was a catalyst for the genomic era allowing for direct sequencing of the genome. In Sanger Sequencing, DNA is denatured into single strands, serving as the template for DNA synthesis. The DNA is synthesized via chain terminating polymerase chain reaction (PCR) that occurs with a mixture containing normal and chain terminating bases (Sanger, Nicklen, and Coulson 1977). As the DNA is synthesized, chain terminating bases are randomly incorporated into the growing DNA strand, creating different sizes of DNA fragments. The fragments are then size-separated via capillary electrophoresis. As fragments exit the capillary, a detector excites the fluorescent labels associated with the nucleotides on the DNA molecule. This information is compiled into a chromatogram with the peaks corresponding to the appropriate nucleotide, returning information about the order of the associated nucleotide at each position (Heather and Chain 2016; Sanger, Nicklen, and Coulson 1977).

This technique of direct DNA sequencing spearheaded the efforts related to the Human Genome Project that began in 1990, taking 13 years to complete and costing $3 billion (Hood and Rowen 2013). Given the extended timeframe of the Human Genome Project, focus was directed to developing technologies that would expedite direct DNA sequencing (Hood and Rowen 2013). This led to the second wave of sequencing technology termed "Next-generation sequencing" (**Figure 1.3**, "Next-generation sequencing") (Goodwin, McPherson, and McCombie 2016). This era of next-generation sequencing was driven by technology created by the company Solexa, which would later be acquired by Illumina (Slatko, Gardner, and Ausubel 2018). In this method of sequencing, RNA is extracted from the sample of interest and converted into cDNA to make a library of sequences of the original RNA. The library sequences become fragmented to create shorter sequences amenable for sequencing. The resulting fragments receive adaptors on both ends of the fragments allowing for identification of individual samples and for the molecules to bind to the flow cell where the RNA-sequencing reaction will occur. The novelty of this next-generation sequencing approach was framed by the introduction of the bridge amplification technique, creating dense clusters of identical DNA molecules within a flow cell, increasing throughput from previous Sanger sequencing methods. Next-generation sequencing technologies utilize a sequencing-by-synthesis reaction using the bridge-amplified clusters as a template. As sequencing occurs, fluorescently labeled nucleotides are added to the growing DNA

strand with a camera capturing the fluorescence signal emitted by the individual nucleotide upon incorporation (Bentley et al. 2008). Next-generation sequencing technology enabled collection and identification of on average one million sequences per run rather than a few hundred tediously collected sequences with Sanger technology (Kukurba and Montgomery 2015).

Bainbridge and colleagues reported one of the first methods for generating transcript sequences utilizing the bridge-amplification approach and Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) search methods to align the collected sequences to their associated gene in the genome (Bainbridge et al. 2006). The term RNA-sequencing was introduced around 2008, describing the process of mapping and quantifying transcriptomes collected via RNA-sequencing approaches (Mortazavi et al. 2008).

Eventually, this process of next-generation sequencing would become commercially available via technology including the Ion Torrent and other subsequent Illumina platforms such as NovaSeq and NextSeq (**Figure 1.3**, "Next-generation sequencing**"**) (Slatko, Gardner, and Ausubel 2018). The reads generated via these platforms spanned 50-500 base pairs (bp) fragments requiring transcript assembly or alignment to map these small fragments back onto the complex genome to determine the gene for which the sequence represents and its frequency (Amarasinghe et al. 2020). The depth of reads needed to adequately represent and quantify the genes expressed in a eukaryotic transcriptome can vary greatly, but the consensus remains that about 100 million reads are appropriate to map and accurately quantify a eukaryotic transcriptome (Conesa et al. 2016). This technology revolutionized RNA sequencing platforms enabling comprehensive transcriptome profiling, providing new avenues for downstream analysis and differential expression platforms. Next-generation sequencing was followed by third-generation long-read RNA-sequencing techniques offered by Oxford Nanopore (ONT) and Pacific Biosciences (PacBio) (**Figure 1.3**, "Third-generation sequencing"). PacBio long-read RNA-sequencing is a focus of this thesis.

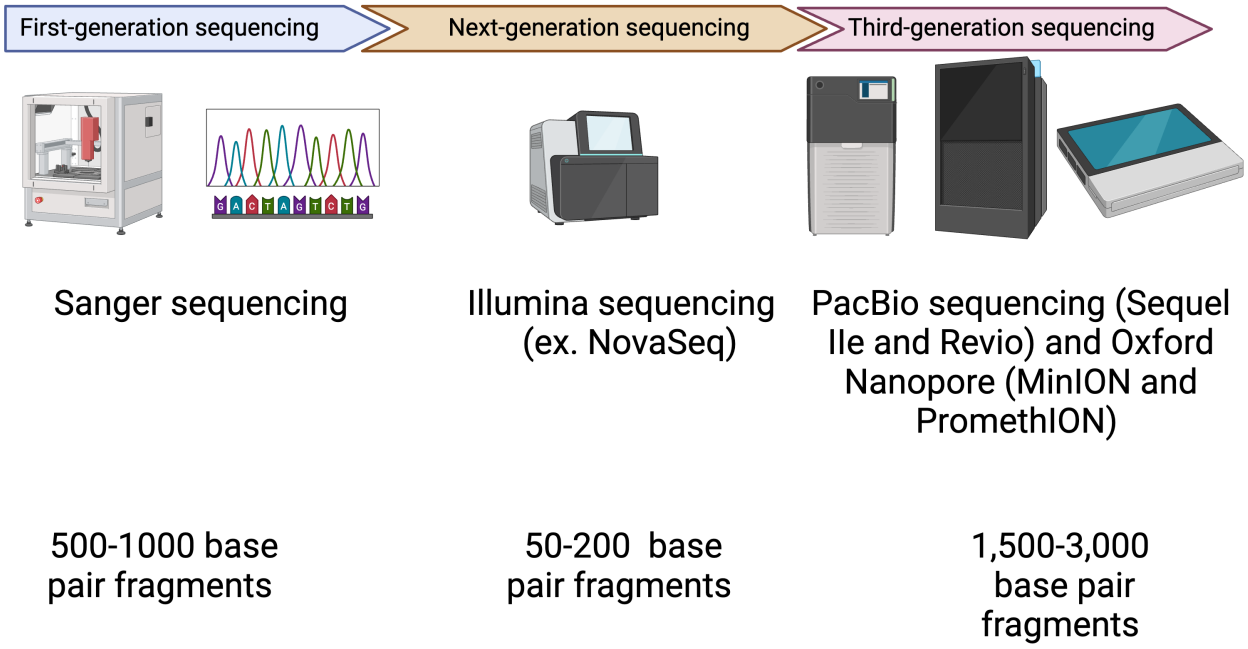


Figure 1.3 Evolution of DNA sequencing platforms and technologies
Illustrating the three-generations of RNA-sequencing technologies. First-generation sequencing, marked by Sanger sequencing technology, utilizes chain-termination methods to sequence DNA fragments. However, while this method was highly accurate, it suffered from low-throughput (Goodwin, McPherson, and McCombie 2016). The second generation of sequencing includes the advent of Illumina sequencing platforms, increasing throughput via technological advancements including parallel reactions and bridge-amplification to accurately sequence entire libraries, producing short-reads around 150 bp in length. While highly accurate, short-reads introduced challenges in reconstructing full-length transcripts for complex regions of the genome. Third-generation RNA-sequencing platforms offered by PacBio and Oxford Nanopore addressed these limitations by producing full-length transcript reads capable of spanning 10kbp. Figure adapted from PacBio (PacBio 2020).

1.2.2 Detecting and quantifying isoforms via RNA sequencing

The starting material begins with RNA from the sample of interest that is converted into cDNA to represent the library of sequences for a sample. Given that cDNA molecules input into sequencing reactions undergo fragmentation, RNA-sequencing reads generally span the length of about 150 base pairs (bp) (Eisenstein 2023). These short transcripts represent small stretches of the transcriptome from the sample of interest that need to be aligned to a reference genome to determine the location for which that transcript belongs. Those reads are then quantified to determine the abundance of that transcript. Thus, this process provides a measurement for a gene's expression level within a sample.

1.2.3 Software for processing short-read RNA sequencing data

Software has been created to process RNA-sequencing data. Specific tools have been created to address splice-aware alignment needs such as HISAT (D. Kim, Langmead, and Salzberg 2015), STAR (Dobin et al. 2013), TopHat (Trapnell, Pachter, and Salzberg 2009), and

Bowtie (**Table 1.1**). This category of splice aware aligners bioinformatically split collected reads aligning to different exons to identify junction-spanning reads that detect alternatively spliced transcripts.

Another category of RNA-software includes transcript reconstruction as accomplished via tools such as CuffLinks (Trapnell et al. 2010) and StringTie (Pertea et al. 2015) which assemble aligned reads into transcript models. Additional software such as Kallisto (Bray et al. 2016), Salmon (Patro et al. 2017), Sailfish (Patro, Mount, and Kingsford 2014), and Bambu (Chen et al. 2023) (with more focus on the benefits of this approach discussed in **Chapter 5**) have been developed to address alignment-free quantification providing computationally efficient estimates of transcript abundance given a reference genome, with each of the tools having their own unique focus (**Table 1.1**). Both transcript reconstruction and alignment free analysis can quantify alternative isoforms.

With the knowledge that specific splicing patterns are associated with particular cell states, software has been developed to measure differential splicing between conditions. The tool Mixture of Isoforms (MISO) focuses on identifying the differentially expressed isoforms between two conditions (Katz et al. 2010). One of the most widely used tools for differential splice analysis is Multivariate Analysis of Transcript Splicing (rMATS) (Shen et al. 2014), which operates by calculating a percent spliced in (PSI) metric to identify a biological condition where a splicing pattern (e.g. exon inclusion) is more predominate in one state compared with its paired condition. rMATs characterizes the five major splicing patterns including exon skipping, alternative 3’ splice sites, alternative 5’ splice sites, and mutually exclusive exons. Another tool, DEXSeq (Anders and Huber 2010) employs a different approach to measure splicing changes by evaluating exons to determine if an exon is statistically more or less used within a transcript when compared across biological conditions. Additional software has also been created such as LeafCutter which characterizes splicing in a reference independent manner by detecting splice clusters related splice variants (Y. I. Li et al. 2018). The tool SUPPA (Alamancos et al. 2015), leverages already prepared transcript quantification results, calculating PSI values of individual splice events between conditions based on transcript expression data.

Purpose	Tool	Features	Citation
Alignment of RNA-sequencing data	STAR	Supports splice-aware alignment and discovery of novel splice sites	Dobin et al. *Bioinformatics. 2013*
	HISAT	Supports splice-aware alignment with inclusion of a annotation file	Kim et al. *Nature.* 2015
	TopHat	Can operate without previous knowledge of known splice sites	Trapnell et al. Bioinformatics. 2009
	Bowtie	Combines speed, efficency and flexible applications for downstream tools and workflows	Lamgmead et al. *Bioinformatics. 2009*
Assembly	StringTie	Allows for the identification of novel transcripts	Pertea et al. *Nature Biotechnology. 2025*
	Cufflinks	Includes assembly, estimation of abundances and differential expression	Trapnell et al. 2010, *Nature Biotechnology*
Quntification	RSEM	Does not require a reference genome, and utilizes an expectation maximization algorithm to rapidly model read distributions	Li et al. *BMC Bioinformatics. 2011*
	Kallisto	Utilizes pseudo-alignment, focusing on the incorporation of k-mers to determime the reads a transcript best matches	Bray et al. *Nature Biotechnology. 2016*
	Salmon	Utilizes a pseudo-alignment approach but incorporates additional metrics for handling biases with RNA-sequencing data such as GC content	Patro et al. *Nature Methods. 2017*
	Bambu	Utilizes a novel discovery threshold to classify novel transcripts or those that are artifcats from sequencing, and can support quantification of for multiple samples	Chen et al. *Nature. 2013*
	Sailfish	Also utilizes a pseudo-alignment approach but focuses on speed and memory	Patro et al. Nature. 2014
Measuring splicing	rMATs	Focuses on calculating a percent spliced in metric (PSI) to identify splicing patterns differing between two conditions	Shen et al. *PNAS. 2014*
	MISO	Compares differential isoform usage between two conditions	Katz et al. *Nature Methods. 2011*
	SUPPA	Relies on quantified RNA-sequencing data utilizing one of the quantification tools above to determine splicing changes between samples	Alamancos et a. RNA. 2015
	DEXSeq	Profiles alternative splicing events by focusing on changes in the exons between transcripts for the same gene	Anders et al. Genome Research .2012

Table 1.1: Overview of software packages for processing RNA-sequencing data with the focus on profiling isoforms

However, such splicing aware platforms are limited by the resolution of the RNA-sequencing derived data.

1.2.4 Limitations of short-read RNA sequencing data

It has been observed that 95% of human genes encompass transcripts that are greater than 300 bp, with the average length of a transcript being 1,712 bp (Li et al. 2019). With most short-read RNA-sequencing platforms returning reads significantly shorter than the length of an average transcript (Kovaka et al. 2019), incomplete transcript read coverage compounds issues of resolving transcript sequences. Alternatively spliced transcripts often combine their sequences in distinct or novel ways, with distinct AS events occurring far apart beyond the length of short-reads (Anvar et al. 2018). Therefore short-reads cannot effectively distinguish alternatively spliced transcripts alone (Au et al. 2013). With the results of RNA sequencing reflecting gene expression dynamics in a cell, inadequate measurements can fail to capture isoform complexity.

1.2.5 Addressing limitations of short-read RNA sequencing for the purposes of isoform detection with long-read RNA-sequencing technology

To overcome the limitations of short-read RNA sequencing for the purposes of isoform discovery, breakthroughs in technology have led to a "third-generation of sequencing", yielding long-read RNA sequencing technologies such as those from PacBio and Oxford Nanopore (refer to **Figure 1.3**, "Third-generation sequencing"). PacBio long-read sequencing technologies can produce reads spanning tens of kilobases (kbp) in length rather than short-read RNA-sequencing technologies, which have a limit of 50-200 bp, with full-length reads eliminating the need for complex assembly and reconstruction approaches (Sharon et al. 2013) (**Figure 1.4**).

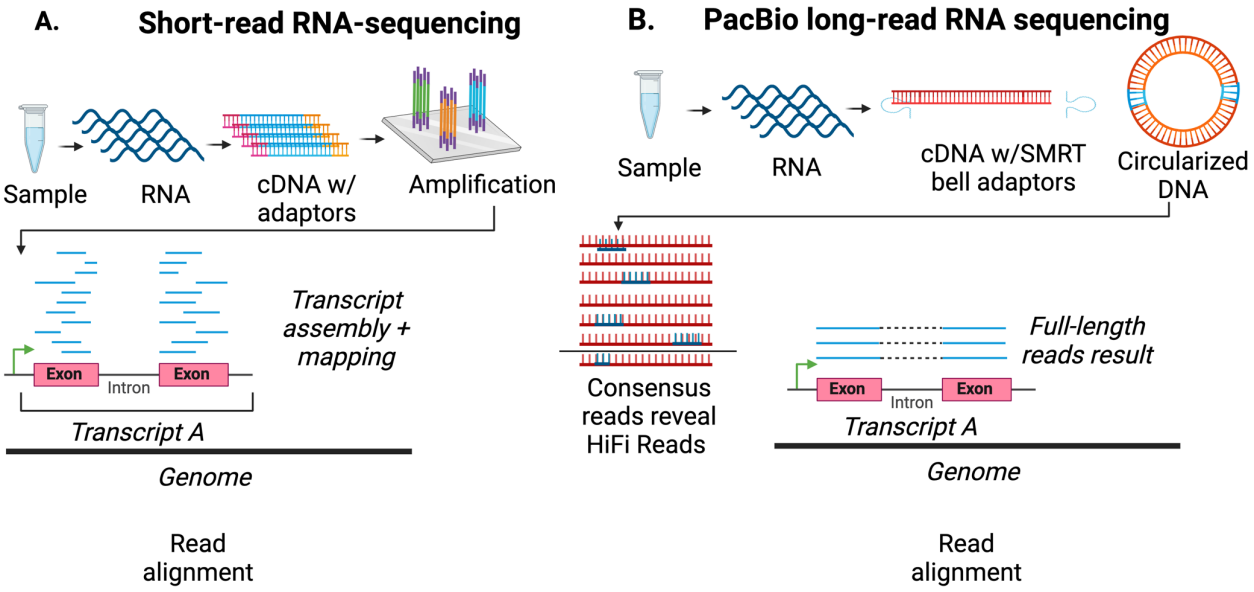


Figure 1.4 Short vs. long-read RNA sequencing approaches
In short read RNA-sequencing (A.); short transcripts are collected then aligned to the genome, with aligned reads used to estimate transcript abundance and reconstruct the transcriptome. Unambiguous mapping of reads to isoforms can occur. In long-read RNA sequencing (B.); reads provide full-length resolution. Reads mapped to full-length transcripts can be counted to quantify isoforms for the respective sample.

The first long-read sequencing technology was termed single-molecule real time (SMRT) sequencing released by PacBio in 2011 (Eid et al. 2009) (**Figure 1.3**, "Third-generation sequencing"). The starting material begins with RNA extracted from the sample of interest which is converted into cDNA that then is extended with SMRTbell adaptors attaching to the blunt ends of the cDNA molecules to form a circularized DNA sequence (Eid et al. 2009). The fragmentation step necessary in short-read RNA-sequencing is omitted to allow for direct sequencing of full-length cDNA molecules. The process of sequencing full-length transcripts from the 5'end to the poly-A site is termed Iso-Seq as developed by PacBio (Ardui et al. 2018).

Single circularized molecules enter the zero-mode waveguide (ZMW), the site of the sequencing reaction. A polymerase is added, orbiting around this circular DNA, synthesizing DNA during each revolution. As nucleotides are added in the growing complimentary strand, fluorescently labeled nucleotides are incorporated within the growing strand, a laser and camera capture the incorporation in real-time. The resulting read, constructed after several revolutions of the polymerase, is a HiFi read constructed by averaging the repeat cDNA sequences collected. The fully constructed reads are on average 15-20 kb in size (Hon et al. 2020), generally able to cover the average length of human transcripts. This process has been widely adopted and carried out through the production of the Sequel sequencing platform by PacBio.

With PacBio sequencing resulting in full-length transcripts, the exact exon-to-exon boundaries are measured, allowing for unambiguous identification of a transcript sequence. Thus, given the full-length resolution of collected transcripts, this technology overcomes limitations of short-read RNA-sequencing, enabling discovery of novel exon-exon connections that may have been difficult to resolve. Additional details on PacBio long-read RNA-sequencing analysis platforms and technology advancements can be found in **Chapter 5**.

1.3 Tracking the protein isoforms expressed in a biological system

1.3.1 Mass-spectrometry based proteomics for detection of protein isoforms

While RNA-sequencing technologies have provided greater resolution of the transcriptome, profiling the associated protein-level product is essential to assess the extent of functional diversity achieved via AS events (Reixachs-Solé and Eyras 2022). Collecting such information allows direct association of an isoform's expression to an associated biological state and potential effect.

To detect the proteins expressed in a sample, mass-spectrometry based proteomics remains the gold standard (Mann et al. 2013). In this application, proteins are extracted from a sample of interest and digested into peptides through enzymatic digests (e.g. trypsin), cleaving proteins at predictable amino acid (AA) locations. The resulting peptides are amenable for bottom-up mass-spectrometry applications, in which the peptides ultimately serve as the proxy to identify a protein product (Nesvizhskii and Aebersold 2005).

After enzymatic digest, the resulting peptide mixture is very complex consisting of tens of thousands of peptides (Shishkova, Hebert, and Coon 2016). To simplify the complex mixture, high-performance liquid chromatography (HPLC) is interfaced to the mass-spectrometer to separate peptides by their hydrophobicity (chemical composition), simplifying the mixture that is then electro-sprayed into the MS instrument.

The simplified mixture of peptides then enters the mass-analyzer in the form of ions using high voltage gradients to allow the peptides in the gas phase to be transmitted through the instrument towards an ion detector. In this thesis, I will be focusing on the mass-spectrometry approach using a ThermoFisher Orbitrap Eclipse system.

In an MS measurement, the ions are collected in the ion trap and are separated by their mass-to-charge ratios (m/z). The instrument performs an initial survey scan called MS1, in which all ions from a user-defined m/z range are measured. Operating under Data Dependent Acquisition (DDA) parameters, a selection of the most abundant ions from the MS1 are then subjected to a second scan (MS2) during which individual precursor ions are subjected to collisions with neutral gas causing the peptide to fragment into product ions. With amino acids discernable by their respective mass-to-charge ratios, such information supports identification of peptides. Over thousands of iterations of MS1 and subsequent MS2 scans, this process can characterize the amino acid sequences of the peptides within the sample using statistical search workflows such as target-decoy analysis (Elias and Gygi 2007). These peptides sequences are utilized to map back to a protein sequence and provide protein-level evidence for the existence of a protein within a sample **(Figure 1.5)**.

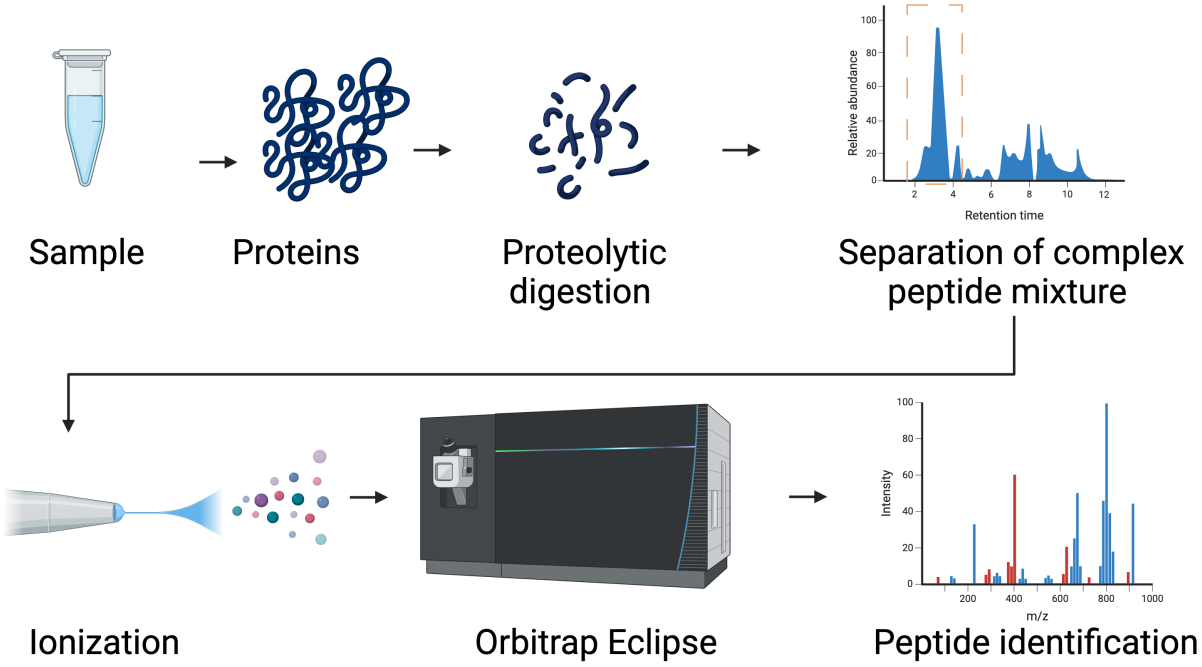


Figure 1.5 Overview of bottom-up mass-spectrometry-based proteomics
The general mass-spectrometry workflow to discover proteins represented within a sample of interest.

One can find a limitation of bottom-up MS analogous to short-read RNA-sequencing, in which the peptides resulting from MS analysis are far shorter than their source protein (often less than 25 AA), therefore unambiguously mapping them back to the original full-length protein sequence of interest introduces ambiguity. Since the proteins are cleaved at specific AAs, this process may not reveal peptides that are uniquely distinguishing for proteins, adding complexity in resolving closely related protein isoforms (Wang et al. 2018).

A limitation of MS-based proteomics is the challenge of detecting peptides for isoform-specific regions. For instances of mRNAs with several isoforms, the resulting protein sequences

for each isoform may have short stretches of AAs that are isoform-unique (or distinguishing), while the vast majority of AA sequence might be shared among different isoforms (see **Chapter 2 Figure 2.1** "Long-read Proteogenomics pipeline") (Miller et al. 2022). There are additional challenges to confirming isoform-unique peptides, including peptide abundance, peptide ionization efficiency, and other "proteotypicity" characteristics of the AA region of interest (Mallick et al. 2007).

Due to the infrequent detection rates of isoform-specific regions, some researchers have concluded that most human genes likely have a single protein isoform (Tress, Abascal, and Valencia 2017). To account for the possibility of true protein isoform expression, other approaches have settled at utilizing an alternative isoform's longest sequence as the reference sequence to increase the chance of identifying peptide sequences that match to some region of that gene during the mass-spectrometry alignment process (Tress, Abascal, and Valencia 2017b). Other tools such as TRIFID and MANE utilize the large amounts of proteomics data available to attempt to classify the importance of splice isoforms (Pozo et al. 2021, 2022).

Uncertainty remains in the field over the importance and the prevalence of factors driving protein isoform complexity within a tissue or sample. Tools such as APPRIS (Rodriguez et al. 2018), recognize the landscapes of at least some isoforms that exist but limit the space by defining a "major isoform", based on several metrics such as conserved domains and other evolutionarily conserved functional elements. However, this approach ignores the fact that tissues exhibit specific expression patterns, and genes can express two or more isoforms across distinct tissues (Baralle and Giudice 2017; Tapial et al. 2017).

1.4 Importance of isoforms in regulating distinct biological processes

1.4.1 Isoforms involved in cell differentiation processes

While studies remain limited in experimentally detecting isoform-specific peptides to support protein-level evidence of protein isoforms, other studies have begun to illuminate the functional importance of isoforms in controlling cellular processes. For example, during differentiation it has been observed that the transcription factor *OCT4,* a master regulator of pluripotency, undergoes AS producing *OCT4A* (expressed during pluripotency) and *OCT4B* (expressed after the induction of differentiation) (Atlasi et al. 2008). Another example is the forkhead box transcription factor (*FOXP1),* where AS events cause an exon skipping event that alters residues in the DNA binding domain affecting the balance between pluripotency and differentiation pathways (Gabut et al. 2011).

AS events do not just regulate the balance of differentiation and pluripotency but can govern cellular activities in mature tissue types. In the process of hematopoiesis, the establishment of new blood cells, the transcription factor *Ikaros* undergoes AS to produce isoforms restricting differentiation pathways of cells (Molnár and Georgopoulos 1994).

1.4.2 Isoforms involved in the pathogenesis of diseases

As mentioned previously, AS events have also been linked to the progression of disease. For example, aberrant AS of the *BCL-x* gene alters the ratio of the healthy *BCL-X_L* isoform that blocks apoptosis, preventing healthy cell turnover, and the apoptotic *BCL-X_S* isoform. This switch in isoform ratios serves as an early indicator for some cancers (Dou et al. 2021; Aguzzoli Heberle et al. 2024).

1.4.3 The importance of isoforms in endothelial cell identity

Of specific focus in this thesis is the expression of isoforms in endothelial cells (ECs). ECs represent an important cell type within the body comprising the luminal lining of all blood vessels allowing for the cardiovascular system to carry out critical functions, such as regulating gas exchange, nutrient delivery, and maintenance of vascular tone (Cleaver and Melton 2003; Godo and Shimokawa 2017). ECs acquire specialized phenotypes to become arteries, capillaries, and veins to dynamically respond to and accomplish their diverse roles within the cardiovascular system (Marziano, Genet, and Hirschi 2021). Due to the importance of these ECs within the cardiovascular system, it is not surprising that dysfunctions arising within these cell types present a myriad of cardiovascular diseases (Sun et al. 2019). Thus, cataloging the factors contributing to the maintenance of EC function and plasticity could yield massive therapeutic potential towards identification of target genes to treat cardiovascular diseases.

It has been observed that ECs can be modulated by the process of AS and express distinct isoforms. In ECs, the gene *Nova2* undergoes AS that governs the organization and expansion of the vascular lumen, which is critical during the process of angiogenesis (Giampietro et al. 2015). One of the most drastic examples showing the functional activities among isoforms is demonstrated in the AS of the vascular endothelial growth factor A (*VEGFA*) in which there have been 13 isoforms identified with their splicing patterns falling into two distinct families; one family of isoforms promotes angiogenesis while the other hinders angiogenic processes. These families of *VEGF-A* isoforms must be in balance within ECs to control growth and expansion of vascular cells (Farrokh et al. 2015).

Another example of isoforms modulating EC function is seen in the endothelial nitric oxide synthesis (*eNOS*) gene with three known splice variants. AS creates novel exons that cause early truncation of the protein, altering polyadenylation signals causing variations in interactions with target genes modulating *eNOS* production and expression (Farrokh et al. 2015). AS also occurs in the gene Endoglin (*ENG*), a transmembrane glycoprotein that expresses multiple isoforms with different affinities for target genes, modifying interactions between pathways (Farrokh et al. 2015).

While these studies have begun to highlight the critical role of isoforms with distinct functions in ECs, the full repertoire of isoforms in ECs remains an unknown. Additionally, providing protein-level evidence and linking such transcript isoforms to protein function remains an area of active investigation. Thus, cataloging the isoform expression profiles underlying EC

identity represents an opportunity to begin to delineate the isoform factors contributing to function.

1.5 Integrating short-read RNA sequencing with mass-spectrometry based proteomics to profile protein isoforms and determine their role in transcriptome complexity

Widespread understanding of the repertoire of isoforms expressed in a system is still an area of active investigation. To better define the exact transcriptome and corresponding proteome content of a cell or tissue, "proteogenomics" approaches can be used. These approaches involve integration of transcriptomics information with matched MS-derived data to link transcripts to their translated protein isoform counterparts (Nesvizhskii 2014; Reixachs-Solé and Eyras 2022).

Early proteogenomic approaches have relied on short-read RNA sequencing which presents limitations in the ability to confidently catalog isoforms due to the fact that isoform-specific sequences may be lost during the transcript reconstruction process, obscuring regions that were pivotal for isoform characterization (Sheynkman et al. 2016; Alexey I. Nesvizhskii 2014). A major technical challenge is obtaining sensitivity to accurately detect short stretches of isoform-distinguishing (referred to as "isoform-informative") regions (Blakeley et al. 2010).

Leveraging advancements in technology to empirically define transcriptomes and proteomes has been essential in cataloging and defining global, high-resolution, and tissue-specific isoform expression.

1.6 Purpose of this thesis

Isoforms are involved in supporting a wide range of functions affecting a multitude of cell states both within diseased and healthy tissues. While we know that over 95% of genes undergo AS, the atlas of isoform expression within ECs remains limited. Insight into the EC isoform landscape can provide the ability to distinguish factors underlying EC heterogeneity activities within the cardiovascular system. As advancements in both next-generation sequencing and MS-based proteomics evolve, clarity into the intricacies of how the human genome outputs a diverse landscape of transcriptional and proteomic products have become apparent. In this thesis I aim to address limitations in our capabilities to measure isoform expression within ECs, including during dynamic processes, such as differentiation. I address this gap in knowledge by integrating high-throughput long-read RNA-sequencing and MS-based proteomics.

In **Chapter 2**, I will discuss the establishment of an integrated multi-omics approach utilizing long-read RNA sequencing and matched mass-spectrometry data to generate custom databases by which isoforms can be profiled in a sample-specific manner (**Chapter 2**). I apply this approach to illuminate the isoform landscape for a well-studied vascular cell type, human umbilical vein endothelial cells (HUVECs) (**Chapter 3**). I then review the known roles of isoforms in globally regulating early developmental processes and their contribution within

regulatory networks working to define cell fate (**Chapter 4**). Given the prevalence of isoforms in governing developmental processes, I decided to characterize isoform expression as they pertain to EC differentiation. I utilized advancements in long-read RNA sequencing technology in the form of an order of magnitude increase in throughput to carry out time course studies (**Chapter 5**). I profiled isoform dynamics across the continuum of primordial EC differentiation, highlighting EC related genes with switches in their isoform expression (**Chapter 6**). And lastly, I summarize the ongoing efforts and growing knowledge enabled through transcriptomic technologies and provide perspectives on the future of isoform discovery and characterization in a precision medicine context (**Chapter 7**).

Chapter 2 Development of the long-read proteogenomics pipeline

This chapter is adapted from:

Miller, R.M., Jordan, B.T., **Mehlferber, M.M**. Jeffery E.D., Chatzipantsiou C., Kaur S., Millikin R.J., Dai Y., Tiberi S., Castaldi P.J., Shortreed M.R., Luckey C.J., Conesa A., Smith L.M., Deslattes Mays A., Sheynkman G.M. Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol* **23**, 69 (2022). https://doi.org/10.1186/s13059-022-02624-y

2.1 Introduction

This chapter summarizes the key findings surrounding the development of the long-read proteogenomic approach and rationale for development. As mentioned in **Chapter 1**, an approach that enables visualization of isoform expression in a sample-specific manner is essential to begin to define the isoforms comprising a tissue specific sample. Such information can support endeavors to understand how isoforms contribute to a phenotype, as proteins form the operational units of the cell. Capturing AS events comprising the transcriptome and the proteome derived from the human genome is important for many facets of biology including characterizing and targeting biologically relevant pathways to support biomedicine and developmental biology studies.

This long-read proteogenomic approach described here, will become the underlying method utilized in the subsequent chapters to support a comprehensive elucidation of the isoform landscape within ECs. Here, and in the following chapters we will expand upon the usage of this approach for uncovering isoform expression in vascular-related cell types to demonstrate the utility of the approach and the knowledge available when utilizing a matched tissue-specific database for isoform characterization.

2.1.1 Uncovering the isoform population comprising a sample

Gaining an understanding of the isoforms comprising tissues in both healthy and diseased states provides vital knowledge of the distinguishing features allowing a tissue or cell-type to achieve its distinct functions. Multiple protein isoforms can arise from the same gene through the process of AS creating distinct arrangements of AA sequences. Protein isoforms resulting from this process can exhibit different stabilities and functional effects (Yang et al. 2016). It has been observed that protein isoforms are involved in a wide range of diseases. However, limited experimental knowledge exists to detect protein isoforms at high resolution, leaving ambiguity to the extent by which transcript isoforms contribute to the complexity of the proteome (Blencowe 2017).

For the process of defining the proteins comprising a sample, MS based proteomics has become the leading method for sensitive characterization of distinct protein populations of the proteome.

To measure the proteome, proteins from a sample are digested into short peptide sequences comprised of AA subsets from the original intact protein. The complex mixture of peptides is resolved through liquid chromatography (LC-MS) methods, separating peptides by their hydrophobicity for analysis. Once separated, these peptides become ionized and enter the mass spectrometer, where they are sorted based on their mass-to-charge ratio (m/z). Next, peptide precursor ions are trapped in the analyzer region, fragmented during collisions with gas atoms, and are then analyzed by the detector. This is the MS2 scan, containing fragment ions that correspond to the amino acid sequence of the precursor ion. Peptide sequences are identified by searching experimental MS2 spectra and correlating them to the predicted fragment ions of known peptides in a protein database. These identified peptides are then mapped back to their potential parent protein based on the alignment of collected sequences. Often, peptides can be shared where their AA sequence matches back to multiple protein isoforms, in those scenarios, uncovering a unique protein isoform based on its sequence alone is limited.

The issue of shared peptide sequences confounds the goal of isoform detection where isoforms often exhibit small and few uniquely distinguishing sequences. Often, these unique regions are not amenable for enzymatic digestion, preventing mass-spectrometry suitable peptides. Therefore, gaining protein level evidence for these unique isoforms at the protein level is extremely difficult. Adding to the complexity of defining protein-level isoform expression, is the fact that peptide identifications are dependent on the contents of the database used. These "reference" databases are often representative of the average tissue expression profiles and can be ignorant towards the variation attributed by tissue-specific phenomena, developmental states, or the differences between diseased and healthy tissues (Mudge and Harrow 2016). Therefore, the composition of the database used for MS analyses directly impacts peptide identifications returned. To optimize the peptide identifications in an experimental design, the protein sequences annotated in the database used for searching ideally should match exactly the sequences represented in a sample. Despite this, obtaining reference databases that are representative of a sample and associated contents is rare. Overall, this discordance can cause a multitude of issues, however for the specific goal of isoform detection in this thesis, the largest obstacle is the inability to identify a uniquely distinguishing peptide sequence, therefore limiting detection of protein isoforms or novel protein isoform events. Such features could be hugely descriptive in defining a cell state or condition.

To address limitations of reference databases, transcript sequencing approaches such as RNA-sequencing can be used to determine expression profiles unique to a sample in order to create a "sample-specific" database that can be used for protein identifications. This database is created based off the transcripts identified via a sample-specific RNA-sequencing experiment

and may be more reflective of the isoform diversity of a sample than the generic reference database.

Efforts to generate a sample-specific database have been achieved primarily through the usage of short-read RNA-sequencing platforms which relies on the short-sequences (as described in **Chapter 1**) to represent a complex genome hindering the ability to effectively delineate complex unique isoform patterns. Long-read RNA sequencing technologies such as those offered through PacBio can delineate full-length transcripts with high fidelity (Frankish et al. 2019). With the full-length transcript resolution, these platforms can uncover thousands of novel isoforms. This presents an opportunity to leverage these capabilities and obtain more complete transcript expression information, which is a prerequisite to equate protein expression and enhance detection of isoform-resolved proteomics. Here, we present the long-read proteogenomics pipeline that achieves enhanced characterization of the protein isoform diversity through the integration of long-read RNA-sequencing with matched MS-based proteomics performed on the same sample. This first-generation pipeline offers the ability to aid in characterizing human protein isoform diversity across various contexts.

Here and in the following chapters, we will expand upon the development of this pipeline. The constructed open-source pipeline and associated modules in this chapter and within **Chapter 3** can be found at https://github.com/sheynkman-lab/Long-Read-Proteogenomics.

2.2 Results

2.2.1 Study Overview

The goal of this pipeline is to provide a sample-specific database that more accurately represents the protein isoform populations expressed in a sample-specific manner and therefore enhance detection of these important isoform populations by leveraging integration of long-read RNA seq and matched MS-based proteomics. The process includes 1. Utilizing and analyzing the PacBio sequencing results to reveal high-quality full-length transcripts 2. Predicting open reading frames from the full-length transcript sequences 3. Utilizing a novel categorization approach, SQANTI Protein to characterize how the transcript sequences are translated to protein 4. Creation of the sample-specific database 5. Profiling novel protein isoforms as determined through the approach. The overall pipeline and associated modules comprising the long-read proteogenomics approach can be found in (**Figure 2.1**).

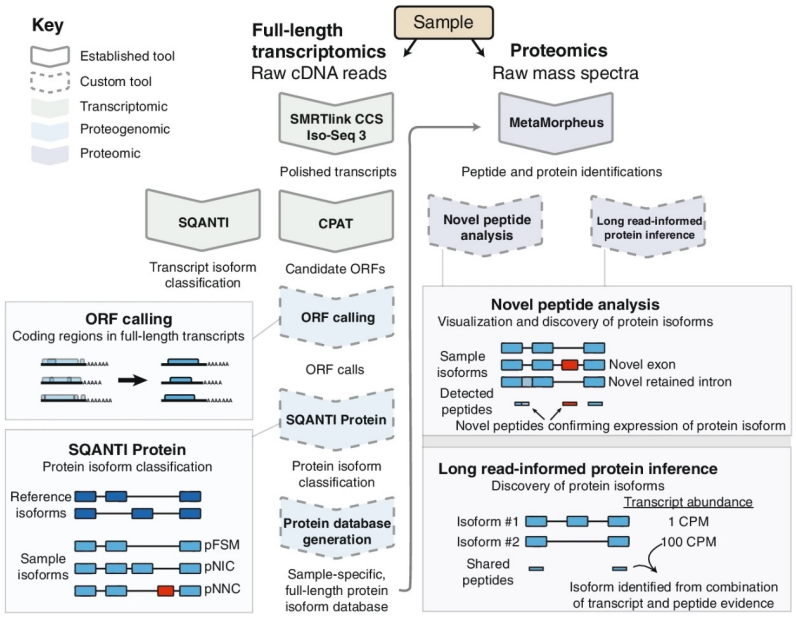


Figure 2.1 Long read proteogenomics approach for enhanced sample-specific protein isoform identification
Schematic of the long-read proteogenomic pipeline for improved protein isoform characterization. The pipeline includes approaches for ORF calling from long read transcripts, an automated protein isoform classification (SQANTI Protein), novel protein isoform detection and a long-read informed protein inference algorithm. CPM – full-length counts per million. Figure made by Rachel Miller.

2.2.2 Long-read RNA sequencing enables identification of isoform populations not represented in the GENCODE annotation alone

Long-read sequencing was performed on the PacBio platform to characterize the landscape of the full-length transcript sequences within the Jurkat T-lymphocyte cell populations. Collected transcript sequences were compared to the GENCODE (A. Frankish et al. 2019) reference database and classified utilizing SQANTI3 (Pardo-Palacios et al. 2023) to determine their associated novelty status. Of the transcripts identified, 43,865 transcripts were known (full-splice matches), 75,491 were determined to be novel **(Figure 2.2A)**. Within that population, 43,075 transcripts were novel-in-catalog having novel combinations of known splice sites/junctions, and 37,416 transcripts were novel-not-in-catalog as they represented unannotated splice sites or exons (**Figure 2.2A**). On average the novel transcripts were in lower abundance than the known transcripts. Overall, the majority of genes expressed multiple isoforms (**Figure 2.2B**). For a third of all genes with observed transcripts, the most abundant protein isoform did not correspond to the "reference" isoform (i.e., GENCODE APPRIS principal reference isoform) (**Figure 2.2C**). Overall, these results highlight the widespread isoform diversity in this sample, emphasizing the need for sensitive methods to define isoform expression.

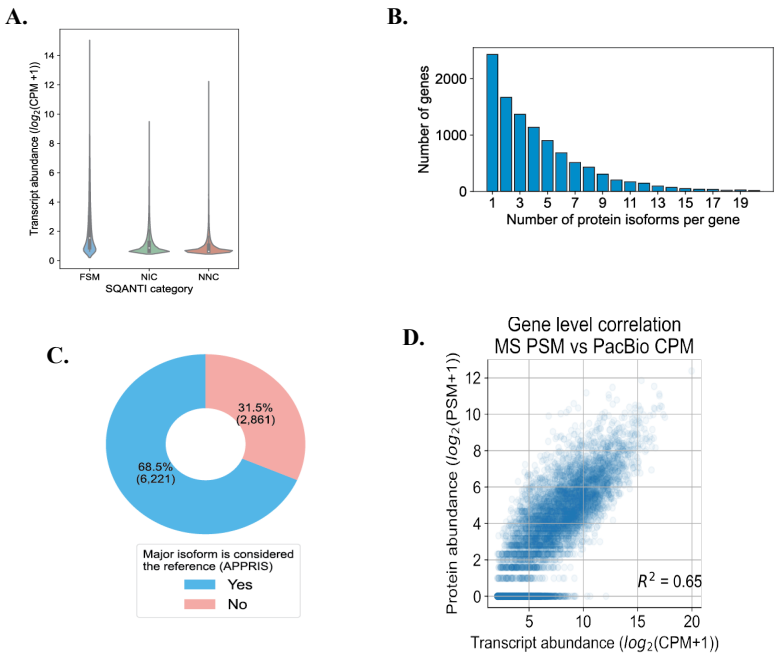


Figure 2.2 Characterization of the predicted candidate protein isoform sequences derived from long-read RNA-sequencing collected data
A. Transcript abundance distributions for the known (full-splice matches (FSM)) versus the novel categories of transcript isoform (novel in catalog (NIC), novel not in catalog (NNC)). B. Bar plot illustrating the distribution of the genes expressing multiple protein isoforms. C. Donut plot illustrating the proportion of genes where the most abundant transcript isoform in Jurkat cells does not align with the identified major isoform as defined by APPRIS. D. Correlation of long-read RNA sequencing derived transcript abundance and protein abundance. Panels made by Rachel Miller and Ben Jordan.

2.2.3 Creation of a classification system for defining and characterizing protein-isoforms

To build a custom sample-specific protein database, we needed a way to derive protein models from the transcripts for each gene. To systematically characterize transcripts into protein models, we crafted a protein classification system extending upon the SQANTI3 transcript classification system to create SQANTI Protein describing the relationship a predicted protein isoform has based on its sequence in reference to GENCODE. This approach relies on the assumption that RNA expression roughly equates to protein expression (Floor and Doudna 2016; Sterne-Weiler et al. 2013). This method enables determination of how a protein sequence model matches the reference and is therefore known (a protein full-splice match, pFSM) or how it may be novel, where elements for that protein contain elements that represent a novel combination (a protein novel in catalog, pNNC) or contain elements that have not been previously annotated (a protein novel not in catalog, pNIC). For a third of the genes with observed transcripts, the most abundant protein isoform for that gene did not match the most dominant isoform as annotated in the reference **(Figure 2.2C)**. This signifies the need for sample sensitive database curation.

It has been previously understood that RNA abundance is at least moderately correlated with protein expression (Y. Liu, Beyer, and Aebersold 2016; D. Wang et al. 2019). We also observed a moderately strong correlation (R-squared = 0.65) when comparing transcript abundance and protein abundance, supporting the use of transcripts as a proxy to represent general protein abundance (**Figure 2.2D**).

After this classification with SQANTI Protein, there were 45,068 protein isoforms (pFSM, pNIC, pNNC) representing 10,348 genes to be considered for the sample-specific database.

2.2.4 Defining a protein database comprised of PacBio derived transcripts

To generate a high-quality database for downstream proteomics analysis, the sample-specific database was crafted in the following ways. Since the database is determined by the collected PacBio transcripts, we found that transcripts with extreme lengths (less than 1kb or longer than 4 kb), having low abundance (less than 3 CPM), or lacking 3' polyadenylation sites were not adequately covered due to the technical limitations of the platform. Any genes that were matching these criteria were excluded and were not considered for the database generation (**Figure 2.3A**). We had relatively high confidence in the models for the genes that passed the filtering steps, and these formed the basis for the high-confidence gene set representing 6,653 genes in the Jurkat cell model. For genes where we were missing an annotation because the high-confidence set did not include that gene, we utilized the GENCODE entries, thus the database would be a hybrid database. This decision ensured that the database would remain comprehensive to maintain integrity for downstream proteomics analysis.

2.2.5 The PacBio-derived sample specific database recovers peptides and genes found in the reference databases

The engineered sample specific hybrid database recovered 99% of the peptide and 99% of the gene identifications that were found when using the GENCODE database alone (**Figure 2.3B**). The overlap of the identified genes and peptides was comparable when using other well-established databases (GENCODE vs UniProt) **(Figure 2.3C)**.

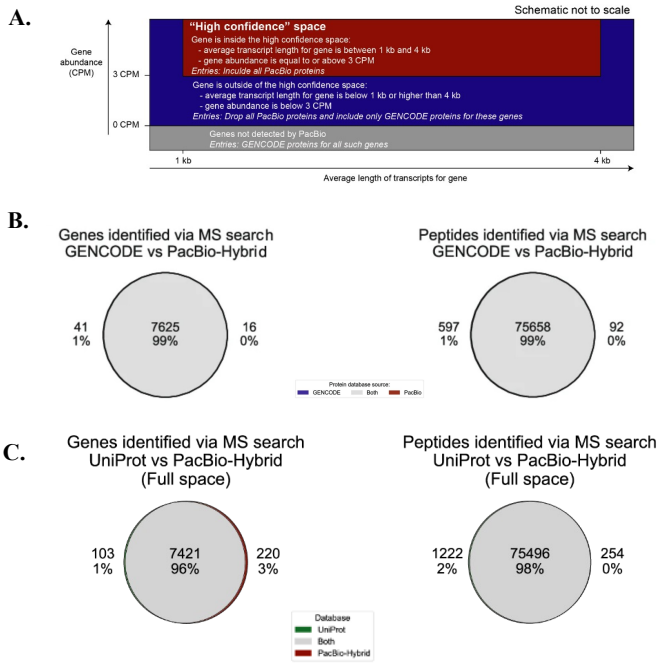


Figure 2.3 Comparison of MS-based proteomics coverage when using different protein databases as the template for MS searching
A. Schematic illustrating the criteria forming selection criteria for the basis of the PacBio-Hybrid database creation where the High-confidence genes are predicted from the transcripts from PacBio long-read RNA-sequencing, where the PacBio-derived transcript read fell short of these parameters, the GENCODE entry is substituted instead therefore making it a hybrid database. B. Overlap of the gene and peptide identifications returned when using the PacBio-Hybrid database or the GENCODE annotation for MS-searching. C. Overlap of the genes and peptides returned when compared against the gene and peptide identifications returned when comparing the PacBio-Hybrid database and the UniProt reference database for MS-searching. Figure panels made by Rachel Miller and Ben T. Jordan

2.2.6 Utilizing the PacBio-derived sample specific database enables identification of novel protein isoforms providing protein level evidence for these cases

Due to the creation of the sample-specific hybrid database, the PacBio-Hybrid database was able to reveal peptide sequences that were not present in either GENCODE or UniProt databases and were therefore novel. In order to verify that these novel peptide sequences were of high quality, stringent criteria were used (generally following the principles outlined in Human Proteome Organization developed Mass Spectrometry Data Interpretation Guidelines (Deutsch et al. 2016)). Upon manual validation of individual novel peptides, we were confidently able to identify 14 peptides which corresponded to unique splicing events and therefore provided protein-level evidence for these splicing events.

Notably, 6 of the 14 novel detected peptides mapped to a single isoform and therefore were unique peptides and provided direct protein-level evidence for the expression of the corresponding full-length novel protein isoform. This direct link is only achievable with the knowledge of the full-length transcripts and the creation of sample-specific sensitivity (Deslattes

Mays et al. 2019). An example of this is outlined in **Figure 2.4** for the peptide ESD, which maps to the PacBio-derived transcript PB.1248.6 and uniquely maps to the region indicating a novel terminal exon. While the prevalence of uniquely mapping peptides is rare, these unique mapped peptides are pivotal components to determine protein-level evidence for protein isoforms.

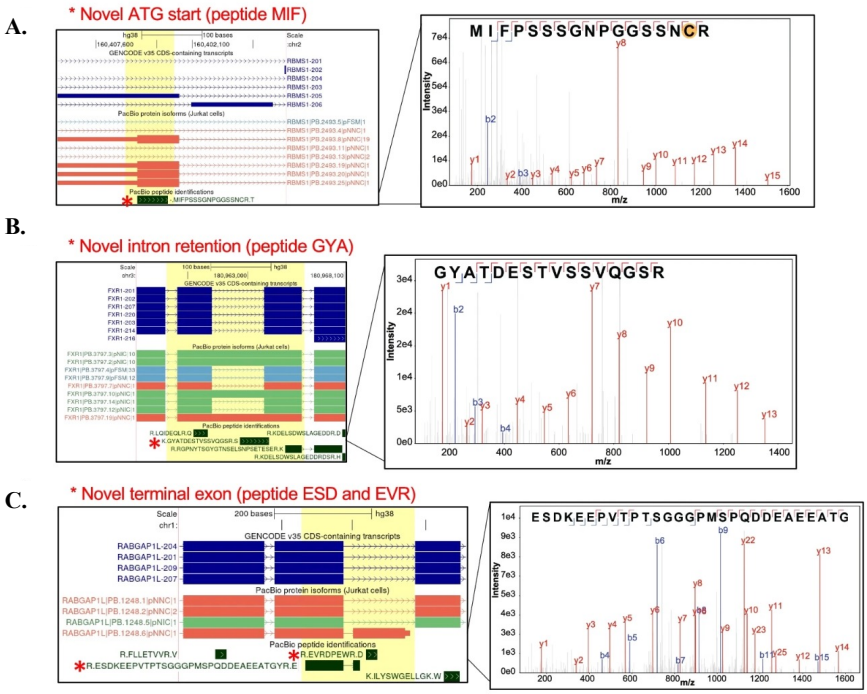


Figure 2.4 Discovery of novel peptides and full-length protein isoforms
A. Novel peptide MIF confirms translation of an ATG start for *RBMS1*. B. Novel peptide GYA confirms translation of a novel retained intron for *FXR1*. C. Novel peptides ESD and EVR confirms the translation of a novel terminal exon for *RABGAP1L*. In this case, since the novel peptide maps exclusively to PB.1248.6, the corresponding full-length protein isoform is likely translated. Note that only ESD passed strict manual annotation, but EVR, which passed a 1% FDR in the global MS search, supports the expression of the same terminal exon.

2.3 Discussion

The overarching goal of this work was to provide a platform enabling the systematic characterization of protein isoform populations, which are the underpinnings for defining healthy and diseased tissue types. To our knowledge, this study represents the first long-read proteogenomics pipeline that integrates and capitalizes on the full-length resolution of full-length transcripts identified via PacBio sequencing in conjunction with matched MS data to accomplish identification of full-length protein isoforms. This work highlights the importance of engineering a sample specific database to enhance protein isoform detection in a sample-specific manner. As sequencing platforms continue to improve accuracy and detection of discrete isoform

populations, the resolution of sample-specific databases derived from transcript sequencing will continue to be more sensitive.

Limitations to this study include minimal experimental approaches that have characterized the individual functions of protein isoforms. While we have established an approach to characterize their expression, further experimental approaches seek to characterize the function and effect of the use of different protein isoforms. Additionally, the space of isoform specific peptides is generally very small, hindering the ability for identification of unique protein isoforms. Deeper MS coverage including usage of multi-protease approaches may work to produce a greater population of isoform specific peptides. Additionally, targeted MS approaches such as Tomahto (Yu et al. 2020) or MaxQuant.Live (Wichmann et al. 2019) that work to target and elucidate specific peptide sequences of interest may support deeper elucidation of distinct and unique protein isoforms produced by alternative splicing.

The flexibility of the pipeline extends to be adaptable towards a wide array of tissue types and samples to interrogate the landscape of isoforms in order to begin characterizing the role of protein isoforms.

The work in this chapter describes the foundational approach for characterizing the landscape of protein isoforms more effectively. This approach will be utilized in the following chapter to elucidate the isoform landscape within vascular cell types.

Chapter 3 Application of the long-read proteogenomics pipeline to characterizing the landscape of human umbilical vein endothelial cells

This chapter is adapted from:

Mehlferber, M. M., Jeffery, E. D., Saquing, J., Jordan, B. T., Sheynkman, L., Murali, M., Genet G., Acharya B.R., Hirschi K.K., Sheynkman G.M. Characterization of protein isoform diversity in human umbilical vein endothelial cells via long-read proteogenomics. *RNA Biology* 2022, 19(1), 1228–1243. https://doi.org/10.1080/15476286.2022.2141938

List of supplementary files associated with this chapter:

- Appendix A: Supplemental Figures (S1-S6)
- Appendix B: Supplemental Tables (S1-S6)

3.1 Introduction

Endothelial cells are critical for the development and maintenance of the cardiovascular system. They form the lining of all blood vessels within the body allowing for functions such as oxygen nutrient delivery, blood pressure regulation, and immune control (Cleaver and Melton 2003). Endothelial dysfunctions can contribute to a host of cardiovascular diseases, such as atherosclerosis, diabetes retinopathy, cancer, and stroke (Rajendran et al. 2013). Improved understanding of these and related diseases may be attained through molecular characterization of the proteome underlying endothelial cell identity and functionality (Richardson et al. 2010; Nordon et al. 2009).

Endothelial cells can express functionally distinct protein isoforms through the process of alternative splicing (AS). For example, vascular endothelial growth factor A (VEGF-A) exists as two separate isoform families that differentially bind to the extracellular region on VEGFR1 or VEGFR2 leading to proliferation and survival of endothelial cells. One VEGF-A isoform family is pro-angiogenic and another is anti-angiogenic (Farrokh et al. 2015; Bowler and Oltean 2019). Together these isoforms work in balance to regulate new vessel formation. Globally, across the endothelial cell proteome, many gene functions are modulated by AS (Mthembu et al. 2017; Di Matteo et al. 2020; Hang et al. 2009; Giampietro et al. 2015). However, despite many high-throughput sequencing datasets collected on endothelial cells (Khan et al. 2019), our knowledge of individual protein isoforms that are expressed is incomplete (Mudge and Harrow 2016).

In order to characterize the proteome of endothelial cells, human umbilical vein endothelial cells (HUVECs) can serve as a relevant model system, since they are primary cells that can be expanded in culture to generate sufficient material for proteomic analysis (Richardson et al.

2010; Caniuguir et al. 2016; Banarjee et al. 2018). A prior study performed by Madugundu and colleagues employed a proteogenomics approach, incorporating RNA-seq and mass-spectrometry (MS)-based proteomics in order to characterize proteomic variation in HUVECs (Madugundu et al. 2019). By utilizing short-read RNA-seq data, the authors generated a set of custom databases of relevance to protein variants. Though the primary focus of the study was to characterize diverse sources of variation, such as single amino acid variants and phosphorylation, they generated a database of candidate splice-junction peptides derived from novel exon-to-exon connections (i.e., junctions), as well as a custom database based on inferred reconstruction of full-length transcripts. The study reported a few novel splice junction peptides, providing further insight into the role of splicing events in HUVECs. However, the proteogenomics approach used relied upon short-read RNA sequencing in the custom database generation, and short reads cannot provide unambiguous knowledge of the bona fide full-length isoform (i.e. complete chain of exon/junction connectivity) (Steijger et al. 2013), which is needed for accurate prediction and detection of full-length protein isoforms (Alexey I. Nesvizhskii 2014).

For improved characterization of protein isoform expression in HUVECs, it would be ideal to obtain full-length transcript information to infer expressed isoforms at the protein level. Fortunately, advances in sequencing technology, such as through the PacBio or Oxford Nanopore long-read sequencing platforms, have allowed for detection of full-length transcript isoforms. Capitalizing on these technologies, we previously developed a proteogenomic approach that incorporates long-read RNA sequencing with MS analysis, which we term “long-read proteogenomics” (Miller et al. 2022). Long-read RNA-seq returns information on full-length transcript isoforms (Sharon et al. 2013), which is bioinformatically translated into full-length protein isoform predictions (Miller et al. 2022; Deslattes Mays et al. 2019; Verbruggen et al. 2021; Anvar et al. 2018). These predicted protein isoforms serve as sample-specific, full-length isoform models from which to infer protein expression from MS data (Alexey I. Nesvizhskii and Aebersold 2005).

Here, we apply a long-read proteogenomic approach to characterize protein isoforms expressed in HUVECs. We demonstrate the application of PacBio long-read RNA-seq data towards characterization of the full-length transcriptome in HUVECs, which includes detection of unannotated transcript isoforms. A PacBio-derived HUVEC protein database is searched against a sample-matched MS dataset facilitating the characterization of HUVEC-specific isoforms. Finally, we report on the discovery of novel peptides, providing evidence for novel isoforms through a direct mapping of novel peptides to full-length protein isoforms in HUVECs. Overall, we present the first application of a long-read proteogenomics approach as applied to primary endothelial cells. These results nominate candidate isoforms for functional studies of how splicing modulates endothelial cell phenotype and function.

3.2 Results

3.2.1 Long-read proteogenomics to characterize isoforms in endothelial cells

In order to characterize the isoforms expressed in an endothelial cell population, we subjected HUVECs to "long-read proteogenomics" where samples undergo long-read RNA-sequencing and mass-spectrometry analysis in parallel, which is followed by integrative analysis of the matched datasets (Miller et al. 2022). The full-length transcripts obtained from long-read RNA-seq are converted to a predicted protein database, serving as candidate isoforms for proteomic detection (**Figure. 3.1**). As a first step in our method, PacBio RNA sequencing is performed to characterize the HUVEC transcriptome.

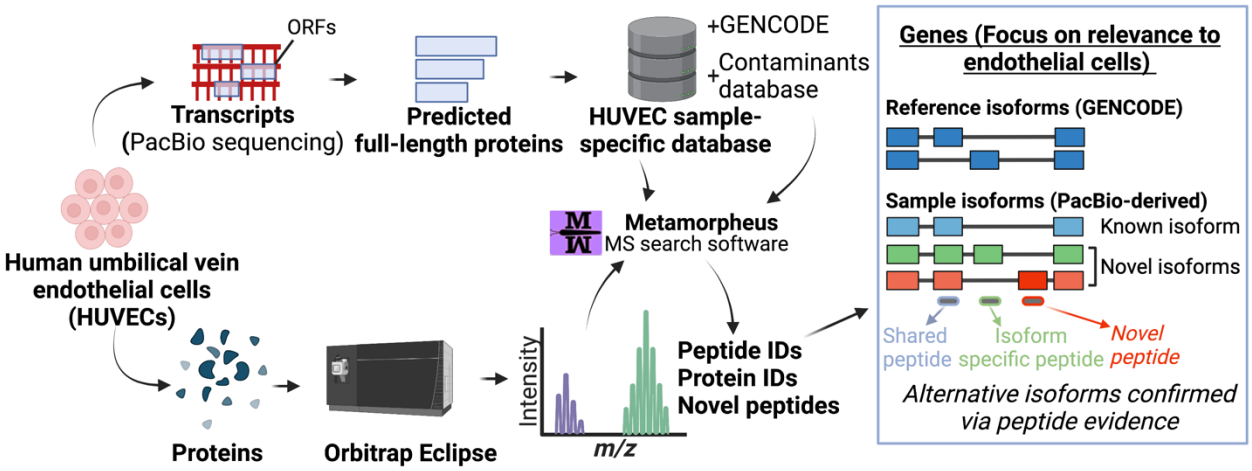


Figure 3.1 Characterization of isoform diversity in HUVECs through integration of long-read RNA-seq with mass-spectrometry data ('long read proteogenomics')
Transcripts are converted into a protein isoform database based on predicted open reading frames (ORFs) and the resulting database is searched against a sample-matched bottom-up mass spectrometry (MS) dataset. The peptide identifications can be used to support the expression of isoform candidates related to endothelial pathways.

3.2.2 Long-read RNA-seq of HUVECs reveals widespread and novel isoform diversity

Long-read RNA-seq data was collected on the PacBio sequencing platform using the "Iso-Seq" method (Gordon et al. 2015), generating 3,608,972 long-reads (i.e. circular consensus reads). These reads were processed by Iso-Seq3 (Gordon et al. 2015) to generate the set of distinct transcript isoforms and their respective abundances (**Figure 3.2A**) (Miller et al. 2022).

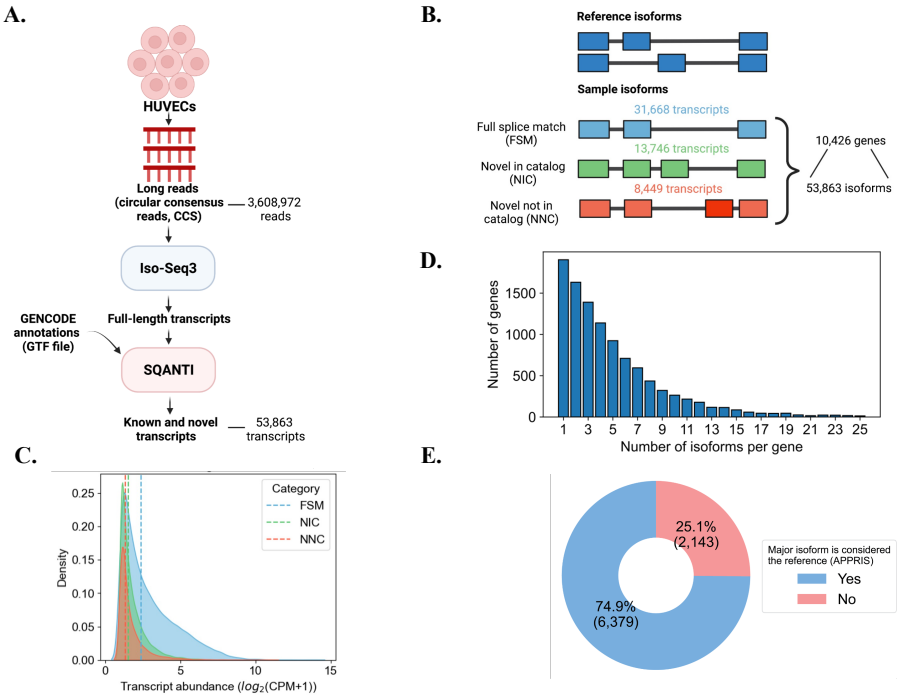


Figure 3.2 Characterization of transcript isoform diversity in HUVECs via long-read RNA-Seq.
(A) Schematic of the long-read RNA-seq analysis pipeline. (B) Transcripts and genes identified from PacBio long-read RNA-seq. The number of known (blue) and novel isoforms (green and orange) are shown. (C) Transcript abundance distribution for known (FSM) versus novel transcripts (NIC, NNC), with dashed lines representing median abundance values in full-length read counts per million (CPM) for each category (FSM = 2.4, NIC = 1.5, NIC = 1.3). (D) Distribution of the number of genes expressing multiple isoforms. (E) Fraction of genes in which the most abundantly expressed isoform ("major isoform") differs from the reference isoform (APPRIS principal isoform).

PacBio-derived transcripts were compared to reference transcripts (GENCODE v35) and their novelty status was defined using SQANTI3 (**Figure 3.2B**) (Frankish et al. 2019; UniProt, Consortium 2019; McGarvey et al. 2019). The UniProt database lacks a complete mapping of protein isoforms to the reference genome, and therefore we could not compare transcripts to UniProt directly, although future efforts may address this limitation (A. Frankish et al. 2019; UniProt, Consortium 2019; McGarvey et al. 2019). Based on a comparison to GENCODE models, we identified 53,863 transcripts from 10,426 protein coding genes, inclusive of all transcripts with a minimum abundance of one full-length read count per million (CPM). The average length of transcripts is 2,846 kilobase pairs (kbp) (*Appendix A: Supplemental Figure S1A*). Among the 53,863 transcripts isoforms identified in the HUVEC sample, 31,668 (59%) matched exactly to a transcript isoform in GENCODE, the match being based on splice junction connectivity ('full splice matches', 'FSM', **Figure 3.2B**). The remaining 22,195 (41%) isoforms were unannotated, or novel, in terms of the observed ordering of splice junctions along the length of the transcript (**Figure 3.2B**). Of the unannotated isoforms identified, 13,746 (62%) contained novel combinations of known splice junctions ('novel in catalog', 'NIC'), and the remaining

8,449 (38%) isoforms contained entirely new exon splice boundaries, in which the acceptor or donor site is not represented in GENCODE ('novel not in catalog', 'NNC', **Figure 3.2B**). The overall abundance distribution for identified transcripts was wide ranging (see *Appendix A: Supplemental Figures S1B*). As expected, on average, the novel transcripts exhibit lower abundance than known transcripts (**Figure 3.2C**) (Huang et al. 2021; Leung et al. 2021). The FSM transcripts displayed a median abundance of 2.4 CPM, while the NIC and NNC transcripts displayed a median of 1.5 and 1.3 CPM, respectively. These data illustrate that novel transcripts tend to exhibit lower abundances than known transcripts. While these trends represent average expression differences, particular novel transcripts can exhibit high abundances within HUVECs.

Using the full-length transcriptomics dataset, we next determined the number of protein-coding genes that returned evidence for expression of multiple isoforms. We found that 82% (8,522 genes of the 10,426 genes represented) of detected genes expressed multiple transcript isoforms (**Figure 3.2D**). To focus on genes involved in endothelial pathways that may be co-expressing multiple isoforms, we manually curated the literature to compile a list of genes that are involved in vascular pathways related to early endothelial differentiation and development or hemogenic specification (*Appendix B: Supplemental Table S1*) (Marcelo, Goldie, and Hirschi 2013; Aragon and Hirschi 2022). We then determined which endothelial genes are expressing multiple isoforms in our HUVEC sample. To have increased confidence in isoform expression of such genes, we filtered for genes which contain two or more isoforms with each isoform having an abundance of at least three CPM. We identified multiple co-expressing isoforms for *CD34*, *CELF1*, *FLT1*, *NRP1* and *SRSF5* (**Table 3.1**, with annotations from GOrilla (Eden et al. 2009; Bowler and Oltean 2019; Lanahan et al. 2013).

Table 3.1 Endothelial-relevant genes expressing multiple transcript isoforms in HUVECs

Gene	PacBio transcript	GENCODE isoform match	Counts per million (CPM)	Function*
CD34	PB.1222.12	CD34-201	13.9	Cell adhesion molecule
	PB.1222.24	novel	9.8	
	PB.1222.26	CD34-202	226.7	
	PB.1222.27	CD34-201	344.7	
	PB.1222.29	novel	12.8	
	PB.1222.31	CD34-203	27.8	
CELF1	PB.7605.14	novel	9.4	Pre-mRNA splicing
	PB.7605.21	novel	4.5	
	PB.7605.23	novel	6.8	

	PB.7605.28	novel	58.3	
	PB.7605.29	CELF1-201	13.2	
	PB.7605.56	novel	5.3	
	PB.7605.70	novel	27.1	
	PB.7605.76	novel	47.0	
	PB.7605.81	novel	10.1	
CDH5	PB.10443.1	CDH5-209	216.5	Regulation of cellular metabolic process
	PB.10443.11	CDH5-201	3.0	
	PB.10443.15	novel	321.8	
	PB.10443.18	novel	3.0	
	PB.10443.2	CDH5-201	2402.9	
	PB.10443.22	CDH5-201	44.0	
	PB.10443.26	novel	9.0	
	PB.10443.28	CDH5-201	13.5	
	PB.10443.33	novel	4.5	
	PB.10443.36	CDH5-201	20.7	
	PB.10443.38	CDH5-201	28.2	
	PB.10443.40	CDH5-208	5.6	
	PB.10443.45	novel	27.1	
	PB.10443.49	novel	9.8	
	PB.10443.50	CDH5-209	6.4	
	PB.10443.52	CDH5-201	10.5	
	PB.10443.57	novel	3.0	
FLT1	PB.8882.15	FLT1-204	42.9	Vascular endothelial growth factor activated receptor activity
	PB.8882.22	FLT1-207	23.7	
	PB.8882.27	FLT1-207	9.4	
	PB.8882.30	FLT1-207	29.7	
	PB.8882.9	FLT1-201	6.4	
NRP1	PB.6952.10	novel	32.0	

	PB.6952.12	novel	3.8	Vascular endothelial growth factor binding
	PB.6952.35	novel	4.1	
	PB.6952.54	novel	3.4	
	PB.6952.58	novel	5.6	
PECAM1	PB.11293.22	novel	29.3	Epithelium development
	PB.11293.23	novel	38.7	
	PB.11293.54	novel	6.0	
	PB.11293.55	novel	5.3	
	PB.11293.64	PECAM1-203	524.8	
	PB.11293.68	novel	32.7	
	PB.11293.7	novel	6.4	
	PB.11293.70	novel	12.8	
	PB.11293.71	novel	4.5	
	PB.11293.80	novel	5.6	
	PB.11293.81	novel	29.3	
	PB.11293.83	novel	3.4	
	PB.11293.9	novel	5.3	
	PB.11293.95	novel	3.8	
	PB.11293.98	novel	3.4	
SRSF5	PB.9356.16	SRSF5-201	10.5	Pre-mRNA splicing
	PB.9356.17	SRSF5-217	103.8	
	PB.9356.21	SRSF5-217	68.4	
	PB.9356.4	SRSF5-207	15.8	

***Function -** GO annotations derived from GOrilla (Eden et al. 2009)

To explore the putative functional effects of candidate genes, we closely examined the potential impacts of changes to amino acid sequences among isoforms of *NRP1*, *CELF1* and *FLT1*. We discovered novel isoforms for *NRP1*, also called neuropilin. *NRP1* is involved in regulating angiogenesis and arteriogenesis pathways through its binding interactions with *VEGF-A* (Marcelo, Goldie, and Hirschi 2013; Lanahan et al. 2013; Kofler and Simons 2015). Notably, we detected a novel isoform (PB.6952.10) at moderate abundance (40 CPM) containing an

alternative donor region. This region has been found within three amino acids of a glycosylation site that has been suggested as potentially affecting neuropilin activity (Bowler and Oltean 2019). Additionally, we found *NRP1* isoform expression of both soluble and membrane-bound forms, and it has been well known that the soluble form acts antagonistically to the full-length form for *VEGF* signaling (Bowler and Oltean 2019). Finally, we identified an isoform with a skipped exon in the C-terminal disordered region of the protein that resides just outside of the transmembrane domain.

Next, we examined *CELF1*, which is an RNA binding protein that is a known regulator of splicing in cardiovascular biology (Chang et al. 2021). We observed eleven isoforms for this gene, seven being novel. The abundance for the major isoform of *CELF1* is moderately high (124 CPM) (PB.7605.2), but the 2nd to 5th isoforms by ranked abundance are also expressed at moderate levels, with two of them novel (PB.7605.5, PB.7605.1). These novel *CELF1* isoforms arise from distinct combinations of splicing events. Nearly all isoforms contain the complete set of three RNA recognition motif (RRM) domains, as described previously (Edwards et al. 2013); however, an alternative acceptor site residing between the 2nd and 3rd RRM domain may alter the inter-domain distance, which may alter binding behavior. Interestingly, the *CELF1* isoforms contain either an extended or truncated N-terminus, which may have a direct effect on cellular localization based on previous reports of the extended N-terminal *CELF1* isoform as being localized to the nucleus and the truncated N-terminal *CELF1* isoform as being localized to the cytoplasm (Blech-Hermoni, Stillwagon, and Ladd 2013). Based on the long-read RNA-seq data, we estimate that in HUVECs, approximately 30% of *CELF1* isoforms may be localized to the nucleus.

Finally, we examined *FLT1*, a gene that encodes the vascular endothelial growth factor receptor 1 (*VEGFR1*) and mediates *VEGF-A* signaling allowing for the survival and proliferation of endothelial cells (Marcelo, Goldie, and Hirschi 2013). We identified ten protein isoforms for *FLT1*. Five of such isoforms were novel and were all extremely low in abundance (~1 CPM, only a few reads supporting their existence); therefore, we did not consider them further. Among all the isoforms, we observed two major families of *FLT1* isoforms: 1) full-length isoforms that contain the transmembrane domain, and can promote endothelial proliferation and angiogenesis (Di Matteo et al. 2020), and 2) short, soluble isoforms that lack the transmembrane domain but still binds to *VEGF-A*, and thus loses its signal transduction function, and therefore is anti-angiogenic (Bowler and Oltean 2019).

Given the prevalence of genes that co-express multiple isoforms in HUVECs, we next asked to what extent the identity of the most highly expressed isoform, i.e. the major isoform, matched what is defined as the "reference isoform" for a gene. To define a gene's "reference isoform," we used the APPRIS database which reports a principal isoform to be most representative for a gene (Rodriguez et al. 2018). The APPRIS principal isoform concept is related to the concept of a UniProt "canonical" protein, though underlying assumptions differ (UniProt, Consortium 2019; Rodriguez et al. 2018). For the genes expressing multiple isoforms,

we classified the corresponding isoforms as either major, i.e. the most abundant isoform based on relative expression levels of all isoforms for a gene, or minor. There were 1,904 genes only expressing one isoform and therefore were excluded from analysis. We identified 8,522 transcripts as the major isoform and 43,437 as minor isoforms. We found, as expected, that on average the major isoforms are more highly expressed than minor isoforms (*Appendix A: Supplemental Figure S1C-D*). Surprisingly, we found that for 25% (2,143 isoforms) of genes, the major isoforms in our HUVEC sample do not coincide with the APPRIS principal isoform (**Figure 3.2E**). Within this population of major isoforms, we found six genes involved in endothelial pathways, *CELF1*, *FLT1, GATA2, NR2F2, NRP1, NRP2* and *SRSF6* (see *Appendix B Supplemental Table S2*). These results illustrate that the major isoform expressed in a given sample may not always correspond to the generic "reference" isoform for a gene, which can be explained by the fact that isoforms exhibit cell or tissue-specific expression patterns (X. Wang et al. 2012).

Next, we examined the presence of previously annotated splice factors (Van Nostrand et al. 2020) expressed within our HUVEC PacBio data. Overall, we detected long reads for 85 annotated splice factors, with the 10 most abundant splice factors including *HNRNPA2B1, HNRNPK, HNRNPC, DDX5, EWSR1, PCBP2, HNRNPA1, PCBP1, FUS, KHDRBS1* (*Appendix B Table S3*). Notably, *SRSF5* was found to be the eleventh highest expressed splice factor at 408 CPM, followed by *SRSF2* as the twenty-second most abundant splice factor at 300 CPM, and lastly CELF1 as the 25th highest expressed at 263 CPMs. Interestingly, it has been observed that SRSF2 and SRSF5 are involved in splicing of VEGF-A pre-mRNA splicing (Farrokh et al. 2015).

We next asked how the novel isoforms differed from the APPRIS principal isoform in terms of length and affected amino acids. As expected, on average, novel isoforms are shorter than the reference form due to the loss of amino acid regions (*Appendix A: Supplemental Figure S2C-D*), with a median shortening of 159 amino acids and an average gain of 11 amino acids. The APPRIS principal isoform for a gene may not be the most representative isoform in HUVECs (see **Figure 3.2E**). Therefore, we also compared the lengths of the novel isoform against the 'major' isoform in HUVECs, i.e. the highest expressed isoform in the HUVEC data. Interestingly, we observed that 'major' isoforms do not tend to be the longest isoform of a gene, or at least this trend is not as stark as with APPRIS principal isoforms. This is likely because the APPRIS algorithm does tend to select the longest isoform of a gene as its 'principal' isoform (Rodriguez et al. 2018).

Collectively, the HUVEC transcriptomic results demonstrate the use of long-read RNA-seq to characterize sample-specific variation in isoform identity and abundance.

3.2.3 Deriving a HUVEC sample-specific protein isoform database

The vast transcriptome diversity of HUVECs likely translates in some part to a diversity of protein isoforms. To explore this question, we translated the HUVEC transcript isoform

sequences in silico into open-reading frames (ORFs) and compiled the predicted sequences into a HUVEC sample-specific protein isoform database for MS searching, as previously described (*Appendix A*: *Supplemental Figure S3A*) (Miller et al. 2022). To classify the relationships between the predicted proteins to that of annotated protein isoforms in GENCODE (A. Frankish et al. 2019), we used the classification scheme we previously developed, SQANTI Protein (Miller et al. 2022). SQANTI Protein automatically categorizes known and novel protein isoforms. The categories include "protein full-splice match" (pFSM), "protein novel in catalog" (pNIC), and "protein novel not in catalog" (pNNC) (*Appendix A: Supplemental Figure S3B*). We found that 16,296 predicted proteins exactly matched protein isoforms in the GENCODE reference (pFSMs), while 24,896 predicted protein isoforms were novel (*Appendix A: Supplemental Figure S3C*). Among those novel isoforms, 5,855 had novel combinations of known protein sequence elements such as the N-terminus, the C-terminus or the splicing pattern (pNICs). The other 19,041 protein isoforms had one or more entirely novel elements, such as a novel N-terminus or an unannotated exon (pNNC).

Among the candidate protein isoforms, we first filtered out protein isoforms that may have resulted from transcripts from incomplete reads or poor-quality transcripts (see Protein database generation in section **3.4 Methods**; 11,876 filtered out). The remaining 34,531 predicted protein isoforms (comprising 16,296 pFSMs, 5,855 pNICs, and 12,389 pNNCs) from 10,912 genes were compiled to create a preliminary HUVEC protein database (**Figure 3.3A**). These genes and their associated isoforms represent candidates for inclusion in the final database. For the final database, we decided to only include isoforms from genes for which we could ensure a complete sampling of the transcripts, and thus the predicted proteins. Therefore, we created a hybrid database in which we defined a core set of genes for which the transcript detection, and thus predicted proteins, is likely complete based on the long-read data collected. The core set of genes included in the hybrid database have a minimum abundance of three CPM and a moderate transcript length (1–4 kbp average GENCODE-annotated transcript length). For all other genes, the hybrid database is populated with all GENCODE protein isoform entries. The hybrid structure of the final database ensures comprehensiveness of the protein models, with the protein completeness assumption of target-decoy searching satisfied so as to avoid issues of an off-target peptide match (Elias and Gygi 2007).

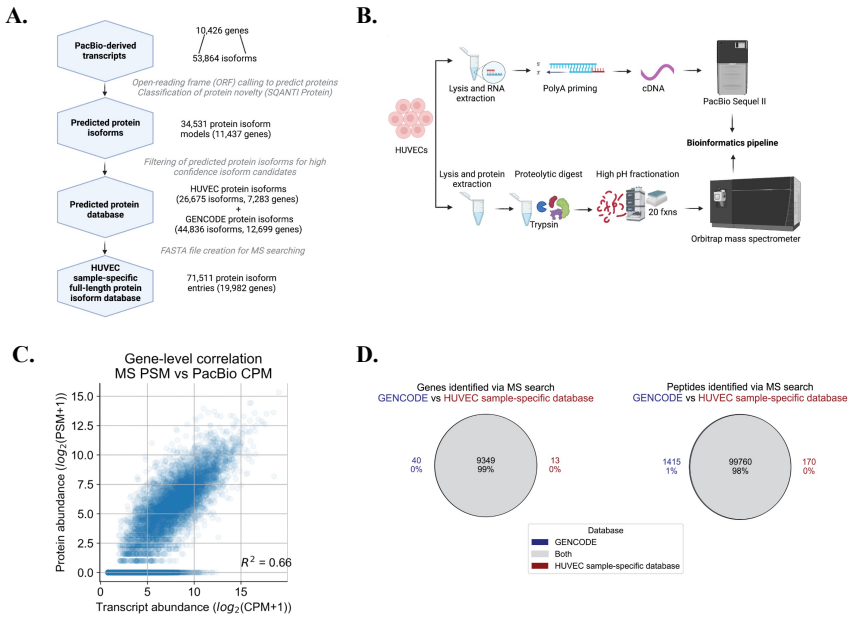


Figure 3.3 Proteomic analysis of HUVECs using a customized long-read-derived protein isoform database

A. Steps involved in the generation of a HUVEC sample-specific database. B. Parallel long-read RNA-seq and MS proteomic data collection from HUVECs. B. Correlation between estimated RNA and protein expression levels. PSM, peptide spectral match; CPM, full-length read counts per million. D. Comparison of proteomic search results between the reference and HUVEC sample-specific database.

As described, the final HUVEC sample-specific database for proteomic analysis includes a mixture of custom PacBio-derived proteins as well as annotated GENCODE proteins (**Table 3.2**). A detailed listing of steps to convert the transcriptome data to a protein database may be found in *Appendix B: Supplemental Table S4*.

PacBio-derived HUVEC sample-specific database		
Source	Genes	Protein entries
GENCODE	12,699	44,836
PacBio-derived (HUVECs)	7,283	26,675
Contaminants (MetaMorpheus)	-	264
Total	19,982	71,511

Table 3.2 Composition of the HUVEC sample-specific database.

3.2.4 Collection of a deep-coverage MS dataset for HUVECs.

To characterize protein isoforms in HUVECs, we generated and analyzed a deep-coverage MS dataset collected on the same HUVEC pellets that were used for long-read RNA sequencing (**Figure 3.3B**). HUVECs were lysed and processed using the filter aided sample preparation (FASP) protocol, in which protein was digested with trypsin to generate a mixture of tryptic peptides. The tryptic digest was subjected to off-line fractionation on an analytical scale high-pH reverse-phase liquid chromatography instrument, and 20 fractions were collected (*Appendix A: Supplemental Figure S4*). These fractions were then analyzed via liquid chromatography LC-MS/MS (Orbitrap Eclipse) in data-dependent acquisition (DDA) mode, generating 3,772,771 MS2 fragmentation spectra. Acquired spectra were searched using the MetaMorpheus (Solntsev et al. 2018) software to obtain peptide and protein identifications passing a 1% false-discovery-rate (FDR). Parameters for the MS search can be found in *Appendix B: Supplemental Table S5*.

3.2.5 The HUVEC-specific protein database returns near-complete coverage of detectable peptides from a reference search

To use PacBio-derived transcripts as the basis for deriving a protein database for MS searching, a key assumption is that the detection of a transcript from PacBio data reflects the discovery of a protein product for that isoform, showing that there is a moderate correlation between transcript and protein abundance. In the past, moderate RNA-protein correlations have been observed using short-read RNA-seq or microarray datasets to quantify transcript abundance (D. Wang et al. 2019; Vogel and Marcotte 2012). Here, we examined the correlation of the transcript abundance that is computed from the long-read RNA-seq data (sum total transcript abundance for a gene, in units of CPM) to the estimated protein abundance (sum total peptide counts passing a 1% FDR, in units of number of peptide spectral matches or PSMs). We observed a moderate correlation with a coefficient of determination (R-square) of 0.66 (**Figure 3.3C**), providing support that the PacBio-based transcript abundances should serve as a reasonable proxy for protein presence, although that may not always be the case for a particular gene.

To assess the general protein sequence content of the HUVEC sample-specific database (not resolved to individual isoforms), we assessed recovery of annotated peptides and genes. The MS data was searched against the GENCODE and UniProt databases to define the set of annotated peptides and genes detectable in the HUVEC sample, and then the same data was searched against the HUVEC sample-specific database. We found that the HUVEC sample-specific database search returned 98% of the peptide and 99% of the gene identifications that were identified when using the GENCODE database for searching (**Figure 3.3D**). The extent of overlap between peptides and genes was similar for the UniProt search results (*Appendix A: Supplemental Figure S5*). Overall, these results indicate that the HUVEC sample-specific database, which was derived de novo from long reads, can capture a majority of the detectable gene and peptide populations likely expressed in HUVECs. Confirmation of the large overlap of

peptide populations identified by the sample-specific database is ultimately useful since it is the underlying populations of peptides identified that are the basis for protein isoform characterization.

3.2.6 Characterization of HUVEC protein isoforms based on available peptide evidence

We have shown that nearly all reference annotated peptides that are detectable are represented in the HUVEC sample-specific database. With the goal of characterizing isoform expression in endothelial cells, we next evaluated the evidence for the presence of isoforms, in terms of the patterns of their underlying peptide identifications. Due to the complexities and potential ambiguities of protein inference (Alexey I. Nesvizhskii and Aebersold 2005), we elected to examine the peptide evidence directly.

We defined three scenarios of isoform detection precision, based on how the set of identified peptides map to isoforms of a gene. The first scenario is when all isoforms of a gene contain only shared peptides, in which the presence of any isoform cannot be definitively confirmed (**Figure 3.4A**, "Protein isoforms correspond to shared peptides"). Among the 10,444 genes with any peptide evidence, we found that 5,993 genes (57%) were cases in which no isoform could be specifically confirmed as expressed because all mapped peptides were shared among two or more isoforms. Of these genes evidenced only by shared peptides, 3,436 are genes containing PacBio-derived protein isoforms in the hybrid database.

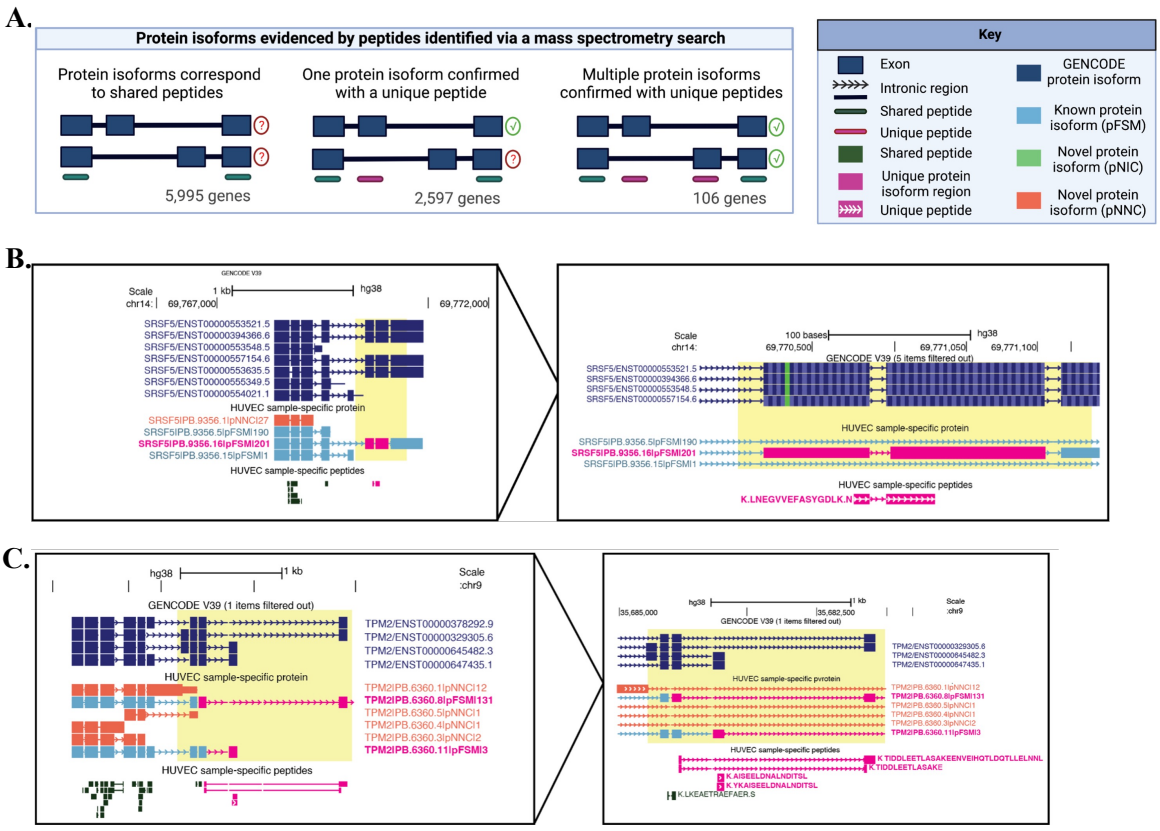


Figure 3.4 Protein isoforms analyzed based on peptides identified via mass-spectrometry
A. Scenarios of differing protein isoform detection precision when evidenced by peptides identified from MS. Only genes with multiple protein isoforms in the database are included, and 1,904 genes that express only one isoform were excluded. B. A protein isoform confirmed with a uniquely mapping peptide LNE, for SRSF5, a splice factor that regulates transcripts of VEGF-A. C. Two protein isoforms of TPM2 are confirmed with uniquely mapping peptides TID, AIS, and YKA. In B and C PacBio-derived protein isoform label follows this format: <Gene>|<PB accession>|<SQANTI Protein class>|<CPM>.

In all other scenarios, there is evidence for the existence of a specific protein isoform because one or more isoforms contain a uniquely mapping peptide. Indeed, the second scenario is when an isoform-specific peptide is identified (**Figure 3.4A**, "One protein isoform confirmed with a unique peptide"). We found 4,451 (42%) genes for which we have unambiguously identified at least one isoform for a gene. For 1,748 (17%) of genes, only a single isoform was listed in the database, thus, all peptides would be expected to be uniquely mapped. For the remaining 2,703 (26%) of genes with multiple isoforms annotated, 2,597 (25%) of genes have a single isoform with unique peptide evidence. For example, we found a single isoform supported by a uniquely mapping peptide (Sequence: LNEGVVEFASYGDLK) for Serine and Arginine Rich Splicing Factor 5 (*SRSF5*), which is involved in the splicing of VEGF-A pre-mRNA (**Figure 3.4B**). Notably, this peptide is shared among two isoforms in the GENCODE database, meaning that the reference database search results cannot pinpoint the source isoform for this peptide.

Of particular interest is a third scenario in which we found evidence for co-expression of two or more isoforms, each supported by a uniquely mapping peptide. In such cases, a natural question is the nature of the functional relationship between the two isoforms and their biological role in endothelial cells. We found 106 (1%) genes with evidence of two or more co-expressing isoforms (**Figure 3.4A**, "Multiple protein isoforms confirmed with unique peptides"). For example, we found two isoforms for Tropomyosin 2 (*TPM2*), each supported by a unique peptide (**Figure 3.4C**). Notably, there were nine genes in which three or more isoforms each had unique peptide evidence. Interestingly, there was an unusually large number of seven protein isoforms detected from the gene Plectin (*PLEC*), which exhibits a series of alternative N-termini due to differential 5' transcription (*Appendix A: Supplemental Figure S6).* A list of all protein isoforms supported by peptide evidence can be found in *Appendix B: Supplemental Table S6*.

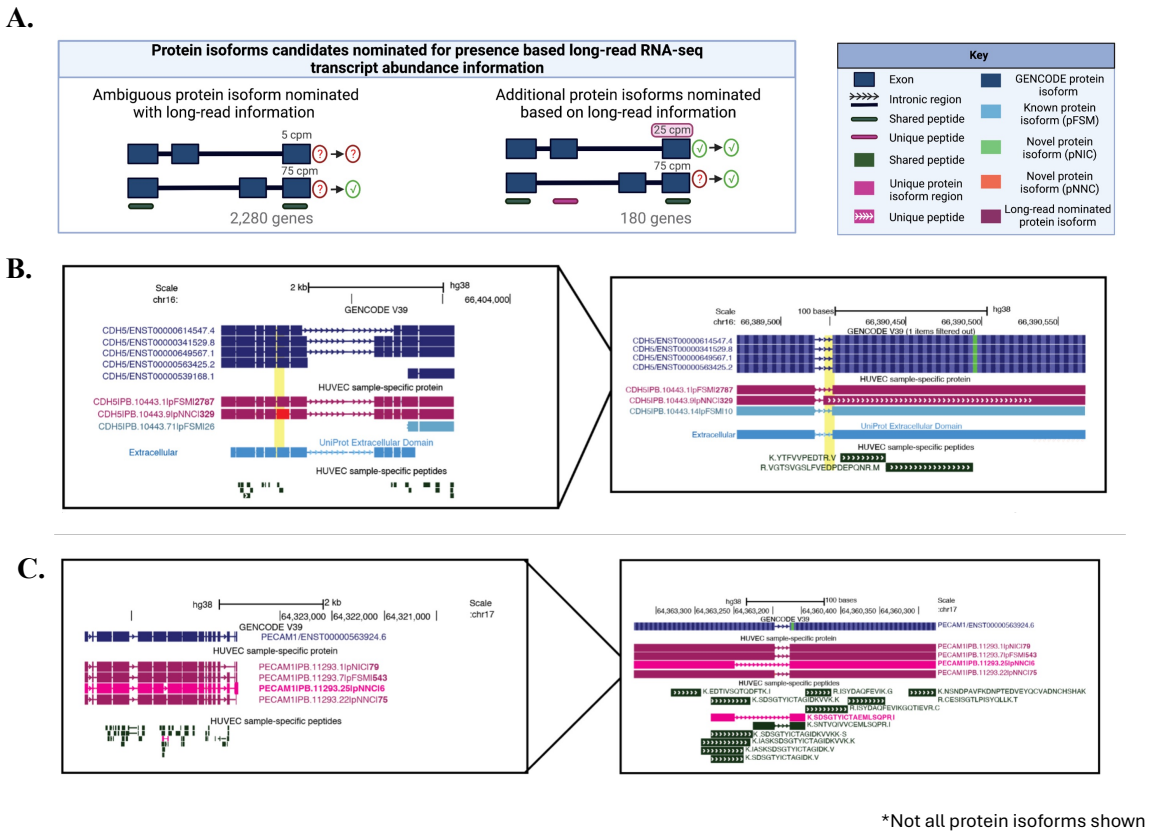
Collectively, these results highlight that while some isoforms may be readily identified from peptide evidence alone, overall, the standard bottom-up MS approach alone does not reach the coverage needed to directly characterize all isoforms predicted from the transcriptome, as observed previously (Blakeley et al. 2010; Lau et al. 2019). Obtaining peptides suitable to resolve protein isoform identification is limited by the peptides detected during bottom-up MS. Part of the challenge is that when comparing isoforms of the same gene, only small stretches of amino acids are unique to an isoform, while the vast majority of amino acid sequence is shared (Miller et al. 2022; Blakeley et al. 2010). Therefore, sampling peptides from the small space of unique amino acids that can directly confirm the presence of a protein isoform is limited by the space of "informative" (i.e. unique-to-an-isoform) peptides (Xiaojing Wang et al. 2018).

3.2.7 Increased support for protein isoform presence in HUVECs through incorporation of underlying transcript evidence from long-read RNA-seq

Despite the use of a sample-specific protein isoform database for MS analysis, a large population of predicted isoforms are only supported by shared peptides (**Figure 3.4A**). Because shared peptides map ambiguously to multiple isoforms, they cannot directly confirm expression of any particular isoform in the sample. However, the evidence for a particular protein isoform could be strengthened by considering the underlying transcript abundance levels provided by the sample-matched long-read RNA-seq data, a concept we previously introduced, as well as described for short-read RNA-seq data (Miller et al. 2022; Carlyle et al. 2018; Salovska et al. 2020). We reasoned that transcript abundance could be used as an additional source of evidence in the isoform discovery process, given there is a moderate correlation between RNA and protein abundance (**Figure 3.3C**).

To explore how long-read RNA-seq data can nominate particular protein isoforms, we first focused on scenarios for which all predicted isoforms for a gene are supported only by shared peptide support. Among such ambiguous protein isoform sets, we reasoned there is higher likelihood for expression of protein isoforms for which the associated transcript abundance is moderately high (e.g., 25 CPM or higher, **Figure 3.5A**). As described in the previous section,

5,993 genes had only shared peptide evidence. Among those genes, 3,436 (57%) contained PacBio-derived isoforms, which have associated transcript abundance information.



*Not all protein isoforms shown

Figure 3.5 Nomination of protein isoforms when incorporating long-read data.
A. Scenarios of protein isoform candidates nominated for expression when transcript abundance from the long-read RNA-seq information is incorporated. B. *CDH5* gene, involved in endothelial pathways demonstrating a scenario of ambiguous protein isoforms identified only by shared peptides, but incorporation of long-read RNA-seq data suggests the expression of three moderately expressed protein isoforms (PB.10443.1, PB.10443.9 and PB.10443.71). C. *PECAM1* gene, involved in endothelial pathways demonstrating an example where one protein isoform is identified via a unique peptide (PB.1123.25), SDS, while the remaining protein isoforms are supported by shared peptides. Abundance information from long-read RNA-seq suggest expression of (PB.11293.1 and PB.11293.7). In B and C, PacBio-derived protein isoform label follows this format: <Gene>|<PB accession>|<SQANTI Protein class>|<CPM>. For B and C, low abundance protein isoforms (<25 CPM) are not shown.

We found that 2,280 (38%) out of the 3,436 genes contain at least one isoform with a moderately high transcript abundance of 25 CPM or higher (**Figure 3.5A**, "Ambiguous protein isoform nominated with long-read information"). Interestingly, we found 247 (4%) genes in which there is potential co-expression of at least two protein isoforms in HUVECs. For example, we found that *CDH5*, otherwise known as VE-Cadherin (Sauteur et al. 2014), potentially expresses multiple protein isoforms. One isoform is highly expressed (PB.10443.1; 2,787 CPM)

and matches a protein isoform in GENCODE and UniProt (GENCODE isoform *CDH5-201*, UniProt accession P33151). However, another isoform is robustly expressed (PB.10443.9, 326 CPM) and, interestingly, is a novel isoform because of alternative usage of a novel splice acceptor (**Figure 3.5B**). This splicing event leads to an isoform of *CDH5* that gains nine amino acids in the extracellular domain, the region of the protein responsible for mediating interactions with other cadherins to regulate endothelial adhesion properties. This example highlights that while *CDH5* isoforms were only supported by shared peptides, the inclusion of the transcript abundance information as provided by the matched long-read data provides higher weights on the existence of at least two isoforms.

To further explore how long-read RNA-seq data can provide additional evidence for expression of protein isoforms, we focused on scenarios in which there is clear evidence for one isoform based on unique peptide evidence, but another isoform of the same gene is supported by only shared peptides (**Figure 3.5A**, "Additional protein isoforms nominated based on long-read information"). We found 180 genes (3%) for which the existence of the alternative protein isoform is supported by long-read evidence (i.e., 25 CPM or higher transcript abundance). Interestingly, we found several protein isoforms of a key endothelial cell surface marker, the platelet endothelial cell adhesion molecule, *PECAM1* (also known as *CD31* (Privratsky and Newman 2014). We found a unique peptide identified for *PECAM1* (PB.11293.25, Sequence: SDSGTYICTAEMLSQPR), but the remainder of peptides identified for *PECAM1* are shared across multiple *PECAM1* isoforms, leaving open uncertainty about the expression of other *PECAM1* isoforms beyond PB.11293.25. From the transcript abundance information, we nominated three additional isoforms accompanied by strong long-read support for *PECAM1* (PB 11293.22, 75 CPM: PB 11293.1, 79 CPM; PB 11293.7, 543 CPM; **Figure 3. 5C**). *PECAM1* produces a transmembrane protein with an extracellular domain, transmembrane-spanning domain, and a C-terminal cytoplasmic domain that likely interacts with intracellular signaling proteins in endothelial cells (Privratsky and Newman 2014; Cao et al. 2002; Dusserre et al. 2004). Strikingly, the differential exon usage observed for these three isoforms are located exclusively in the C-terminal domain, suggesting potential changes to interactions with intracellular signaling molecules. Further details on candidates identified via long-read abundance information can be found in *Appendix B: Supplemental Table S6*.

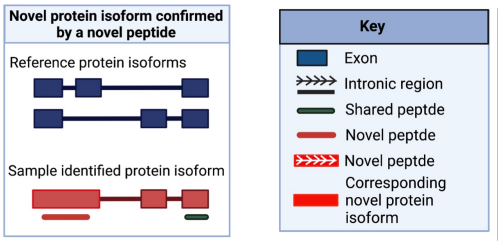
Collectively, these case studies highlight how inclusion of transcript abundance information could nominate protein isoforms which were unable to be directly confirmed as expressed based solely on MS peptide evidence. Note that this approach does not provide any information on the absence of protein isoforms with lower transcript abundance, but, rather, is supplying additional lines of evidence to nominate protein isoforms that may have higher likelihood of expression and represent candidates for functional study. Such isoforms are attractive candidates for further MS validation and subsequent functional analysis.

3.2.8 Novel protein isoform discovery enabled through the HUVEC sample-specific database

We have shown that utilization of a HUVEC sample-specific protein database, with the accompanying transcript abundance values, can lead to inference of novel protein isoform presence. A more direct way to confirm the presence of a novel protein isoform is by detecting a uniquely mapping novel peptide. However, the knowledge of the full-length protein isoforms expressed within a sample is not always possible when using short-read RNA-seq, which can return information on individual splice junctions but may not accurately define full-length transcripts (Frankish et al. 2019; UniProt, Consortium 2019). Long-read RNA-seq provides the full-length transcript and, by extension, the full-length protein isoform prediction; therefore, a novel peptide that directly maps to the full-length protein isoform lends support for its existence.

Using the sample-specific database, we discovered novel peptides for HUVECs, indicating that the reference proteome does not comprehensively capture all protein isoform diversity in a sample. We found 108 novel peptide sequences passing a global 1% FDR, for which they are not represented within the GENCODE or UniProt databases (**Figure 3.6A,** *Appendix B: Supplemental Table S7*) (A. Frankish et al. 2019; UniProt, Consortium 2019). Increased false positive rates for novel peptides have been observed previously (Castellana and Bafna 2010); therefore, we employed strict validation criteria for the novel peptides. Of the 108 novel peptides identified, 39 peptides had a Q-value score below 0.001, corresponding to a 0.1% global FDR. Upon manual annotation of these 39 peptides, we noted 30 peptides with especially strong spectral support, such as full ladders of b- and y- ion fragmentation peaks in the MS2 raw spectra. These novel peptides supported expression of novel alternative transcription or splicing events, such as retained intronic regions or novel exons (**Figure 3.6B,** *Appendix B: Supplemental Table S7*).

A.



B.

Annotation of the novelty event	Count
Alternative acceptor site (3' splice site)	4
Alternative donor site (5' splice site)	3
Alternative first exon	5
Exon skipping event	7
Frameshift due to exon skipping event (number of nt skipped not divisible by 3)	2
Intron retention event	6
Novel exon	3
Usages of an ATG start that is further upstream than the ATG start annotated in GENCODE	1

C.

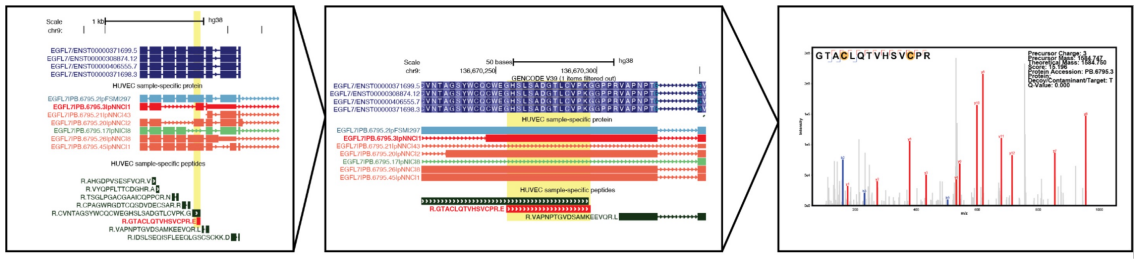


Figure 3.6 Novel protein isoforms discovered via unique peptides.
A. Novel protein isoform confirmed by identified novel peptides. B. Table of the frequency of events supported confirmation of a novel peptide. C. Novel peptide found for a protein isoform of endothelial gene *EGFL7*. Novel peptide and corresponding protein isoform shown in red, which supports a frameshift event for the protein isoform PB.6795.3.

Of the identified novel isoforms, we closely examined splicing events for genes previously implicated in endothelial pathways. Such novel isoforms could represent attractive candidate isoforms for further functional characterization. We found a novel peptide (Sequence: GTACLQTVHSVCPR) confirming the expression of a splice-induced frame-shifted region of *EGFL7* (Protein entry: PB.6795.3), a gene reported through the literature to be involved in vasculogenic pathways as well as hemogenic specification (**Figure 3.6C**) (Nichol and Stuhlmann 2012; Schmidt et al. 2009). We also discovered two novel peptides for *PECAM1*. This is an important finding since *PECAM1* is a marker for endothelial cells and plays a role in the regulation of junctional integrity of endothelial cells and vascular barrier (Privratsky and Newman 2014). Specifically, we discovered a novel peptide (Sequence: ELELLTSKDPPPSASQSAGITDLGKK, maps to protein entry PB.11293.45) corresponding to a novel exon, as well as a second novel peptide (Sequence: SDSGTYICTAEMLSQPR, mapping to protein entry PB.11293.25) that confirms the usage of a novel alternative donor site (*Appendix B: Supplemental Table S6)*.

3.3 Discussion

Endothelial cells that line all blood vessels are critical for the cardiovascular system and their behaviors can be modulated by protein isoforms, though the extent of this mechanism is not known. To characterize isoform expression in endothelial cells, we performed long-read RNA

sequencing (PacBio) of HUVECs to characterize transcript isoforms, and predicted proteins via their translation in silico to protein isoform sequences. To assess evidence for protein isoform expression, we performed MS analysis on the same HUVEC sample and used the HUVEC sample-specific database for MS searching. This general approach has been described and termed "long-read proteogenomics" to enhance protein isoform characterization (Miller et al. 2022).

Our long-read proteogenomics workflow applied to HUVECs, led to the identification of 53,863 distinct transcript isoforms, of which 22,195 were novel. We also found 8,522 genes co-expressing multiple isoforms. Surprisingly, a quarter of the time, the most abundant isoform in HUVECs did not match the predicted "reference isoform" (GENCODE APPRIS principal isoform). This includes genes annotated in endothelial pathways including *CD34* and *NRP1*. From the transcript sequences, we derived a hybrid protein isoform database that contains the highest confidence protein isoform predictions from PacBio-derived transcript isoform sequences, which was completed with GENCODE reference and contaminant sequences. The long-read-derived database captures almost all peptides and proteins detected from searches against the GENCODE protein database.

We identified 10,444 genes with peptide evidence. Based on the peptides identified through MS searching, we found support for expression of 4,451 genes based on uniquely mapping peptides. For the remaining 5,993 genes only evidenced by shared peptides, we incorporated the underlying transcript abundance information as an additional layer of evidence, nominating an additional 2,280 genes as potentially expressed. This group includes a novel isoform for endothelial gene *CDH5* (VE-Cadherin). This case exemplifies how a combination of the full-length transcript and proteomics data can lead to the discovery of novel protein isoforms that cannot be identified by MS data alone. We showed that the HUVEC sample-specific database enabled the discovery of 108 novel protein isoforms based on novel peptide identifications. Among the novel protein isoforms identified is the endothelial gene *PECAM1*.

Our proteogenomic method shows promise for isoform discovery in endothelial cells, but opportunities exist for further improvements. First, limitations in the MS coverage mean that proteins with low abundance or poorly ionizable peptides remain undetected. Future work could involve targeted proteomics, such as parallel reaction monitoring or advanced targeted acquisition strategies, for sensitive detection of alternative protein isoforms (Gallien, Kim, and Domon 2015; Erickson et al. 2017; Wichmann et al., n.d.). Second, the isoforms discovered in this study represent the results of a single cell line in a static culture condition. For the purpose of identifying isoforms that are dynamically regulated, multiple conditions should be examined. Third, the sample-specific database relies on the assumption that sequenced transcripts reflect protein sequences. Thus, we assume that transcripts are both fully sampled as well as moderately correlated to protein expression, which may not be the case for all genes. And finally, our pipeline so far is focused on proteins arising from genes already annotated as protein-coding. An interesting future direction would be to include long non-coding RNAs or other ostensibly non-

coding transcripts, which may reveal coding potential through the proteogenomics approach (Mattick 2018).

Overall, we have shown the application of a long-read proteogenomics platform towards characterization of known and novel isoforms in primary endothelial cells. This approach can uncover isoform populations that could modulate endothelial cell phenotype and function. The systematic discovery of isoforms produces information to guide selection of candidate isoforms for functional studies. This approach can be extended to various endothelial cell contexts including both healthy and diseased states to chart isoforms changing across development or during onset of cardiovascular disease.

3.4 Methods

3.4.1 HUVEC Cell culture

Primary Human Umbilical Vein Endothelial Cells (HUVECs) were purchased from Lonza (C2519AS) and used up to passage five. Early passage HUVECs were cultured in EGM™2-Bulletkit™ medium with growth supplements CC-3156 & CC-4176 purchased from Lonza.

At 80% confluency, HUVECs were trypsinized, washed twice with phosphate-buffered saline (PBS), pelleted, and frozen at −80°C.

3.4.2 Long-read RNA-seq (PacBio Iso-Seq) library preparation and sequencing run

PacBio (Iso-Seq) data were collected on the extracted total RNA collected from the HUVEC cell pellet. HUVEC RNA was analyzed on an Agilent Bioanalyzer to confirm concentration and RNA integrity for downstream analysis. We observed a RIN value of 10. From this RNA, cDNA was synthesized using the NEB Single Cell/Low Input cDNA Synthesis and Amplification Module (New England Biolabs).

Approximately 200 ng of HUVEC cDNA was converted into a SMRTbell library for usage with the Iso-Seq Express Kit SMRTbell Express Template prep kit 2.0 (Pacific Biosciences). Through this protocol, bead-based size selection occurs in order to remove low mass cDNA (less than 500 kb). Each SMRTbell library was sequenced on the SMRT cell on Sequel II system. A 2-hour extension and 3-hour movie collection time was used for data collection. The ‘ccs’ command from the PacBio SMRTLink suite (SMRTLink version 9) was used to convert raw reads into Circular Consensus (CCS) reads.

3.4.3 Mass spectrometry-based proteomics sample preparation

Harvested HUVECs, approximately 5 million cells each, were pelleted and frozen at −80°C. The sample pellet was lysed according to the Filter Aided Sample Preparation (FASP) protocol (Wiśniewski 2018). Lysis buffer used in the FASP was changed to 6% SDS, 150 mM DTT, 75 mM Tris-HCl. To the 30 µL pellet of 5 million cells, an aliquot of 60 µL of lysis buffer was added and probe-sonicated to lyse the cells and shear the nucleotide material. Sonication

continued for 1–5 minutes until the sample was clear and no longer viscous. The lysate was then incubated at 95°C for 5 minutes. Protein quantitation was estimated by BCA assay to be approximately 500–600 µg. Quadruplicate aliquots of 20 µL each were subjected to FASP and trypsin digest (1 µg per aliquot) and allowed to incubate at 37°C overnight. Nanodrop analysis estimated peptide content at 22 µg per trypsin digest (total of 88 µg).

3.4.4 Offline HPLC Fractionation

The tryptic digests were pooled and dried down to a volume of 40 µL and subjected to offline high pH RP-HPLC fractionation using an Agilent 1200 HPLC. Sample was loaded onto a ThermoFisher Scientific Hypersil Gold C18 column (150 mm × 3 mm × 3 µm C18), equilibrated with 95% solvent A (20 mM NH4 formate, pH 10) and 5% solvent B (70% acetonitrile/30% solvent A), and eluted at a flow rate of 400 µL/min, with fractions collected every 1 minute from RT 38–63 min. The following gradient was used: 5% B from 0 to 30 min, 5–65% B from 30 to 63 min, 65–100% B from 64 to 69 min, 100–5% B from 69 to 70 min, 5% B from 70 to 73 min. Samples containing peptide, according to UV 214 nm corresponding to the HUVEC pellet were digested with trypsin. Collected fractions 4–20 were selected for LC-MS/MS analysis.

3.4.5 NanoLC-MS/MS analysis

The resulting peptides were dried to 12 µL and analyzed by nanoLC-MS/MS using a Dionex Ultimate 3000 (Thermo Fisher Scientific, Bremen, Germany) coupled to an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Three microliters of each peptide-containing sample were loaded onto an Acclaim PepMap 100 trap column (300 µm × 5 mm × 5 µm C18) and gradient-eluted from an Acclaim PepMap 100 analytical column (75 µm × 25 cm, 3 µm C18) equilibrated in 96% solvent A (0.1% formic acid in water) and 4% solvent B (80% acetonitrile in 0.1% formic acid). The peptides were eluted at 300 nL/min using the following gradient: 4% B from 0 to 5 min, 4–28% B from 5 to 210 min, 28–40% B from 210 to 240 min, 40–95% B from 240 to 250 min and 95% B from 250 to 260 min.

The Orbitrap Eclipse was operated in positive ion mode with 1.9 kV at the spray source, RF lens at 30% and data dependent MS/MS acquisition with XCalibur version 4.3.73.11. Positive ion Full MS scans were acquired in the Orbitrap from 375 to 1500 m/z with 120,000 resolution. Data dependent selection of precursor ions was performed in Cycle Time mode, with 3 seconds in between Master Scans, using an intensity threshold of 2×10^4 ion counts and applying dynamic exclusion (n = 1 scans within 30 seconds for an exclusion duration of 60 seconds and ± 10 ppm mass tolerance). Monoisotopic peak determination was applied and charge states 2–6 were included for HCD scans (quadrupole isolation mode; 1.6 m/z isolation window). The resulting fragments were detected in the Orbitrap at 15,000 resolution with standard AGC target.

3.4.6 Long-read RNA-seq analysis, MS searching, and proteogenomic analysis conducted using a Nextflow pipeline

The long-read proteogenomics pipeline was implemented with Nextflow, a workflow framework which allows for scalable and reproducible computational analysis. The Nextflow pipeline developed and described previously was used to process HUVEC collected PacBio data, translate the resulting transcripts into the protein database (see Deriving a HUVEC sample-specific protein isoform database below), and perform proteomics database searches (Miller et al. 2022). Further information on the workflow including individual modules of the Nextflow pipeline can be found at https://github.com/sheynkman-lab/Long-Read-Proteogenomics (Miller et al. 2022). The GitHub revision (i.e., commit) used in this analysis was https://github.com/sheynkman-lab/Long-Read-Proteogenomics/releases/tag/v1.0.0. All transcriptomic and proteogenomic docker images that are used within the analysis can be found at https://hub.docker.com/ under the repository gsheynkmanlab. The analysis was performed on the University of Virginia High Performance Computing system.

3.4.7 Long-read RNA-seq (PacBio Iso-Seq) data analysis

The CCS reads obtained from PacBio sequencing were analyzed using the Iso-Seq workflow described previously (Miller et al. 2022). Primer removal was done on the 5' and 3' end. The 5' primer consists of an NEB adapter sequence (Sequence: GCAATGAAGTCGCAGGGTTGGG). The 3' primer consists of the Clontech SMARTer cDNA universal primer (Sequence: GTACTCTGCGTTGATACCACTGCTT). Following processing of the raw reads using the Iso-Seq workflow, we derived the number of full-length reads corresponding to each distinct transcript. Full-length read counts per million (CPM) were computed by dividing the number of full-length reads aligning to a transcript isoform by the total number of reads and then multiplying this by a factor of 1,000,000.

3.4.8 Transcript isoform classification and filtering

SQANTI is a computational tool used for comparison, classification, and quality assessment of the full-length isoform sequences collected from the long-read platform (Tardaguila et al. 2017). We used SQANTI3 (version 1.3) to annotate the polished transcript isoforms obtained from the Iso-Seq analysis using SQANTI default parameters. Note: the default parameters included options to use the genome-derived sequences for the isoform output. As a result, transcriptional variations inclusive of alternative N-termini, alternative splicing, etc. but not genetic variations are captured in the HUVEC sample-specific database.

3.4.9 Generation of a full-length protein isoform database from the long-read RNA-seq data

After deriving a high confidence set of full-length transcript isoforms within the Nextflow pipeline, we select the most biologically plausible ORF for each of the Iso-Seq transcripts. Calling the best ORF consists of two steps: finding candidate ORFs (50 nucleotides or longer) using CPAT (L. Wang et al. 2013), and selecting the most plausible ORF based on

coding potential, relation of AUG start site to GENCODE reference start sites, and number of AUGs skipped to reach the ORF start site.

To generate the PacBio-derived protein database (HUVEC sample-specific database) employed for downstream MS searching, transcripts were grouped that produced ORFs of the same sequence. The total transcript abundance for each grouping was calculated as the sum of all CPM values for the transcripts comprising that group. Candidate isoforms are further classified based on the protein sequence in relation to the reference protein isoforms, as defined in the "sqanti_protein" and "protein_classification" modules in the Nextflow pipeline. Classifications are based on a variant of nomenclature used within the SQANTI3 software, which we call "SQANTI Protein".

Additional filtering was performed in order to retain only isoforms that were likely protein coding. Isoforms that did not have a stop codon within the predicted ORF, and could represent truncations, were removed. Isoforms that were either fully mapped to a protein-coding GENCODE reference isoform ("protein full splice match", pFSM) were retained, as well as isoforms that contained a novel combination of known splice sites or junctions ("protein novel in catalog", pNIC). Of the isoforms that contain novel splice sites ("protein novel not in catalog", pNNC), suspected nonsense mediated decay (NMD) isoforms were removed. Here, NMD suspects were defined as isoforms that contained more than two junctions after the stop codon. Isoforms that were not classified as pFSM, pNIC, or pNNC were removed from consideration. Protein classification details can be found within the "protein_classification" module of the pipeline, while the filtering criteria can be found within the "protein_filter" module of the Nextflow pipeline.

A hybrid database was developed that incorporated isoforms from PacBio if the gene resided in the high confidence region, defined as where the aggregated transcriptomic gene abundance contained at least three CPM and the average reference transcript length was between 1 and 4 kilobases (kbp). If a gene did not meet these criteria, the reference isoforms were substituted in place of the long-read isoforms. If a gene was not found within the long-read transcriptomic data, the reference protein isoforms were also appended into the hybrid database. A detailed description of reasoning behind creation of a hybrid database has been described previously (Miller et al. 2022).

3.4.10 GENCODE and UniProt reference protein database

The GENCODE protein database used in this study was created by downloading the coding translation FASTA and grouping entries with the same protein sequence for each gene ("make_gencode_database" module in the Nextflow). For the many cases where one or more GENCODE transcripts from the same gene lead to the same protein sequence, the transcripts were grouped and assigned a protein accession as the first alphanumeric GENCODE protein accession, by the transcript name (e.g., GAPDH-201).

The UniProt database used was the reviewed human database with isoforms, downloaded November 1st, 2020. The database contains 42,358 protein isoform entries from 20,292 genes.

3.4.11 MS database search

Standard proteomic analysis of acquired mass spectra files were performed using the free and open-source search software program MetaMorpheus (Solntsev et al. 2018). A custom branch and Docker image were made as part of the Nextflow pipeline (GitHub: https://github.com/smith-chem-wisc/MetaMorpheus/tree/LongReadProteogenomics , Docker: https://hub.docker.com/r/smithchemwisc/metamorpheus/tags?page=1&ordering=last_updated tag: lrproteogenomics) based on MetaMorpheus version 0.0.316. Analysis of the collected spectra files performed either using the HUVEC sample-specific database (HUVEC-derived PacBio reads + GENCODE entries; 'HUVEC sample-specific database') (71,511 of entries from 19,982 genes) in which the subset of PacBio derived entries are 26,675 protein isoforms from 7,283 genes. The GENCODE human database (version 35; 87,729 protein entries from 19,982 genes), or the UniProt reviewed human database with isoforms (downloaded 8 July 2021; 42,380 protein entries from 20,292 genes). All searches were conducted with a contaminants database, included in MetaMorpheus, which contains 264 common contaminant proteins frequently found in MS samples.

All RAW spectra files were first converted to mzML format with MSConvert prior to analysis with MetaMorpheus (see "mass_spec_raw_convert" module in the Nextflow pipeline). For the MetaMorpheus MS search, the settings used for all search tasks can be found in *Appendix B: Supplemental Table S6*. MetaMorpheus produces peptide spectral match (PSM), peptide and protein group result files, which we analyzed in downstream custom modules. All peptide and protein results reported employ a 1% False Discovery Rate (FDR) threshold after target-decoy searching (Elias and Gygi 2007).

3.4.12 Criteria for Novel Peptide Identification

Stringent filtering criteria and manual validation were used, as described previously (Miller et al. 2022; Lloyd M. Smith et al. 2021) to ensure that the spectrum does in fact represent the novel peptide sequence. Spectra corresponding to the scan number of the identified novel peptide sequence were derived from MetaDraw and manually inserted into an Excel file which were then manually evaluated. Corresponding University of California Santa Cruz Genome Browser tracks depicting protein isoforms were derived and can be found via the following session: https://genome.ucsc.edu/s/mm5db/211018_huvec_hcd_trp. In addition to previously (Miller et al. 2022) described criteria for novel peptide annotation, we allowed for cases where the C13 isotope for a novel peptide was selected as the precursor.

3.4.13 Data analysis and plot generation

All downstream data analyses were performed through custom Python scripts. Data analysis scripts used for generation of figures, plots, and statistics may be found in the following GitHub repository: https://github.com/sheynkman-lab/Huvec-Proteogenomic-Analysis

3.4.14 Availability of data and materials

Raw long-read RNA-seq data collected on the PacBio platform are available from the Sequence Read Archive (PRJNA832812, corresponding to accession SRR18959149). Data generated by mass spectrometry are available through MassIVE, the Mass Spectrometry Interactive Virtual Environment (MSV000089326). The output of the data analysis including the long-read proteogenomics Nextflow workflow results generated using the mass spectrometry and long-read RNA-sequencing data as well as the post pipeline analysis results are available on Zenodo (https://zenodo.org/record/7117445#.Y2FQE-wpD0o).

The open-source software produced in the making of this work is freely available under the MIT license found in the GitHub repository (https://github.com/sheynkman-lab/Long-Read-Proteogenomics). A wiki was created (https://github.com/sheynkman-lab/Long-Read-Proteogenomics/wiki) describing each of the pipeline processes.

Code used to generate the main figures and tables in this manuscript can be found in the GitHub repository (https://github.com/sheynkman-lab/Huvec-Proteogenomic-Analysis).

Chapter 4 Importance of isoforms in dynamic settings, the role of isoforms in development with specific focus on isoforms of splice factors and transcription factors

This chapter is adapted from:

Mehlferber, M.M., Kuyumcu-Martinez, M., Miller, C.L., Sheynkman G.M. - Transcription Factors and Splice Factors—Interconnected Regulators of Stem Cell Differentiation. Curr Stem Cell Rep 9, 31–41 (2023) https://doi.org/10.1007/s40778-023-00227-2

4.1 Stem cells as a powerful system for studying development and disease

4.1.1 Stem cells differentiate into diverse cells - Waddington landscape and molecular patterns

Stem cells can differentiate into any cell type in the body. Initially, human embryos were used to generate embryonic stem cells (hESCs) that are derived from the inner cell mass (ICM) of the blastocyst (Thomson et al. 1998). Such cells are pluripotent, with the ability to differentiate into any cell of the body given the proper genetic or exogenous factors. Another source of stem cells are terminally differentiated somatic cells, which can be reprogrammed into induced pluripotent cells (iPSCs) by addition of four key transcription factors (TFs)—*OCT4, SOX2, c-MYC* and *KLF4*, as shown in the landmark experiment in the Yamanaka lab (Takahashi and Yamanaka 2006). iPSCS are largely indistinguishable from embryo-derived hESCs in terms of their genetic, molecular, and phenotypic properties, permitting widespread application of these cells for disease modeling without the ethical issues of human embryo use (Salomonis et al. 2016; Takahashi and Yamanaka 2016).

Stem cells can model the transition from a pluripotent to a differentiated state, which is critical for development of specialized cell types and tissues. Stem cell potency exists in a continuum, with successive cell divisions correlating with narrower differentiation potentials. Such transitions can be thought of as a series of cellular states, which has been analogized by Conrad Waddington as a marble (the “cell”) traveling down a hilly terrain to arrive at a position of the energetically most favorable cellular attractor state (the terminally “differentiated cell”) (Creighton and Waddington 1958). Cell states are reflected by global patterns of gene expression, especially transcript and protein molecular expression. Tracking such coordinated molecular expression changes in these interim steps of differentiation can provide insight into the underlying regulatory network logic associated with these cell states, the relationship of which is

a critical question for the stem cell field (Panina et al. 2020; MacArthur, Ma'ayan, and Lemischka 2009).

4.1.2 Stem cells are tractable models to link molecular variation to development

Stem cell models provide a portal to study otherwise inaccessible aspects of in vivo human development, particularly at the genetic and molecular level. Protocols are now available to direct differentiation of stem cells into hundreds of cell types (Rowe and Daley 2019; Sharma et al. 2020). For example, in cardiovascular development, stem cells can be differentiated into arterial and venous endothelial cells subtypes (Dejana, Hirschi, and Simons 2017; Sriram et al. 2015), and hematopoietic development have well characterized models (Kennedy et al. 2012). And, it has been shown that hESCs can differentiate into epicardial cells that graft onto damaged heart tissue for repair (Bargehr et al. 2019). More recently, going beyond the constraint of monolayer 2D cell cultures, 3D organoid cultures have been developed for a series of organ types that better recapitulate in vivo phenomena such as cell-cell interactions and soluble factor gradients (J. Kim, Koo, and Knoblich 2020).

In addition to hESCs, human iPSCs (hiPSCs) can also be differentiated into several different cell types and are widely used to understand underpinnings of development. For example, in the heart, both hECs and hiPSCs can be differentiated into several different cell types including endothelial cells, endocardial cells, and cardiomyocytes to define molecular drivers of cardiovascular development (Shi et al. 2017).These differentiated cells can then be cocultured to study the communication between these cells that build up the heart. Furthermore, iPSCs obtained from patients are excellent tools to determine the mechanisms responsible for disease pathogenesis as well as identifying developmental defects that give rise to these complex disease phenotypes. Stem cells also allow temporal analyses of molecular and cellular events that occur during development. Genome edited hiPSCs or hESCs carrying patient specific mutations are used to model cell specific defects that give rise to human diseases well as to perform screens of compounds or drugs for treatment of disease complications (Shi et al. 2017). Both hESC and hiPSC mediated stem cell model systems are utilized to identify transcriptional regulatory networks necessary for development and function (Yeo and Ng 2013). With advancements in RNA sequencing and computational methods, post-transcriptional regulatory networks necessary for stem cell differentiation are becoming more appreciated.

Tracing the expression changes of factors associated with differentiation temporally can indicate links between genetic factors and the downstream developmental pathways they regulate (Mahla 2016; Young 2011). Though not fully recapitulating *in vivo* complexity, the journey a stem cell takes during in vitro differentiation at least partially recapitulates molecular changes occurring during development and can provide a tractable experimental system with human relevance. The practical benefits include the ability to culture cells in vitro to generate sufficient material for high-throughput molecular assays and biochemical and genetic screens. The benefit of human relevance arises from the fact that human stem cells should best recapitulate the

repertoire of transcript and protein molecular forms (e.g., isoforms, proteoforms) that are primate or human-specific. Though many genes are conserved between human and model organisms, the molecular details of gene products, such as splicing patterns, tend to diverge greatly (Mouse Genome Sequencing Consortium et al. 2002).

The generation of certain cell types is possible by addition of individual soluble factors or transcriptional regulators (Yeo and Ng 2013; Lesha et al. 2023). In large part this knowledge arose from trial and error or small-scale screens, guided by simple morphogenic or gene expression patterning (Zhou et al. 2008; Yamanaka 2008), rather than a fundamental understanding of the underlying gene regulatory network that governs cell behavior. Even such factors operate within an interconnected and complex gene regulatory network; and, therefore, an incomplete picture remains of the underlying molecular logic and critical factors that direct differentiation of stem cells into fully functional, mature cells (Vierbuchen and Wernig 2012). The field of stem cell systems biology aims to complement focused functional studies, through paradigms that merge high-throughput datasets and computational models.

4.2 Gene regulatory networks - TFs mediate pluripotency and differentiation

One type of network that plays a critical role in cell fate decisions are gene regulatory networks (GRNs). A GRN is represented by the set of active transcription factors (TFs) that bind to and regulate their target genes. TFs comprise about 10% of all protein-coding genes (Lambert et al. 2018). TFs can serve as an activator that promotes transcription of the gene or a repressor that inhibits, and thus lowers gene transcription. Within a network representation, TFs can be modeled as nodes, from which one or more directed edges point to target genes. The edge has a sign, depending on the activating (+) or repressive (-) function of the TF (Bulyk and Walhout 2013).

TFs networks are characterized by biochemical or genome-wide techniques (Farnham 2009). The specific DNA sequences to which a TF binds can be assayed using high-throughput approaches (Lambert et al. 2018) such as systematic evolution of ligands by exponential enrichment (SELEX) (Lambert et al. 2018; Jolma et al. 2010), protein binding microarrays (PBMs) (Berger and Bulyk 2006), or microfluidic chip-based mechanically induced trapping of molecular interactions (MITOMI) (Rockel, Geertz, and Maerkl 2012). Cell-type specific TF binding within a cellular context can be inferred genome-wide using approaches such as DNAse-Seq (Galas and Schmitz 1978) or ATAC-Seq (Buenrostro et al. 2015). TF binding can be directly mapped to genomic sites with approaches such as Chip-Seq (Johnson et al. 2007) or Cut&Run (Skene and Henikoff 2017).

4.2.1 TFs as drivers of stem cell fate - GRN knowledge to guide stem cell engineering

TFs can play an outsize role in influencing stem cell fate by modulating gene expression patterns in differentiating cells (Oh and Jang 2019). To accomplish this in stem cell differentiation, TFs are thought to concurrently repress pluripotent genes to allow activation of

lineage-specific genes (Yeo and Ng 2013). The potent effects of TFs is evident by the fact that *OCT4*, *c-MYC*, *KLF4* and *SOX2* form a core network that can reprogram a somatic cell to an iPSC (Takahashi and Yamanaka 2006). Just one or a few TFs can, solely, convert cells to certain lineages. A quintessential example is *MyoD*, which can convert somatic cells into muscle cells (Davis, Weintraub, and Lassar 1987). Many other lineage defining TFs exist across the differentiation spectrum, such as *ETV2*, which is necessary and sufficient for converting mesodermal precursors to primordial endothelial cells (Aragon and Hirschi 2022).

About 1,564 TFs have been annotated (Ng et al. 2021), but for many their role in driving cell fate decisions remain uncharacterized (A. C. Wilkinson, Nakauchi, and Gottgens 2017). Recently, this has been tested across most human TFs experimentally (Ng et al. 2021). A study employed a library of all human TF open reading frame (ORF) clones to test the role of individual TF expression in differentiating cells (Ng et al. 2021). Surprisingly, 241 of the 1,564 TFs tested resulted in a differentiation phenotype across biological replicates. More recently, an even more comprehensive library of human TFs that included all RefSeq annotated ~3,500 TF isoforms were overexpressed in a population of stem cells and, using single cell RNA-seq as a readout—each TF was profiled in terms of its ability to drive diverse cell-type-specific gene expression programs (Joung et al. 2023). For a subset of TFs, the effect of multiple TFs on cell fate were tested, showing both cooperative and antagonistic relationships and reflecting the combinatorial nature of TF activities.

Cellular gene expression is the product of the timing and location of the collective activities of all TFs in a cell. The correlation between TF activity—most commonly, mRNA levels as a proxy—and expression of target genes, either experimentally or through inference via co-expression, is the basis for inferring links between TFs and their downstream targets (Kinney et al. 2019). These correlations guide computational predictions of TFs responsible for determining cell fates. For example, approaches such as CellNet (Cahan et al. 2014) and Mogrify (Rackham et al. 2016) leverage knowledge of TFs and their downstream targets and pathways responsible for promoting differentiation transitions in order to nominate TFs that would be ideal factors for cell engineering (Kinney et al. 2019; Wells and Choi 2019).

4.3 Regulation beyond transcription - alternative splicing regulatory networks

GRNs and the resulting transcriptional output is significantly affected through the process of alternative splicing (AS) (Stamm et al. 2005). During AS, intronic regions are excised from nascent mRNA while the remaining exon coding regions are ligated together to form distinct mRNA isoforms. Such isoforms can greatly expand the protein functional diversity of the cell (Kelemen et al. 2013; Nilsen and Graveley 2010). The first large-scale RNA-seq datasets of human tissues have revealed that 95% of human genes undergo AS (Pan et al. 2008), with many tissue-specific AS expression patterns (Yang et al. 2016).

4.3.1 Splice regulatory networks - the central role of splicing factors

AS is controlled by the splicing regulatory networks (SRN) operative in a cell. The workhorse of SRNs is the spliceosome, a protein machine comprises over 200 proteins (Wachtel and Manley 2009) and five ribonuclear proteins (U1, U2, U4/U6 and U5) (Lee and Rio 2015; Ule and Blencowe 2019). The process of splicing unfolds via the sequential binding of small nuclear ribonucleoproteins (snRNPs) on the pre-mRNA, which eventually catalyze the joining of particular exon junctions (Wachtel and Manley 2009). The activity of the spliceosome is regulated by a repertoire of splice factors (SFs), a subfamily of approximately 356 (Van Nostrand et al. 2020) RNA binding proteins (RBPs) that bind to sites on pre-mRNA and interacts with components of the spliceosome to enhance or inhibit splicing reactions at certain exons (Fu and Ares 2014; He et al. 2023; Quattrone and Dassi 2019). Families of SF exist, such as heterogeneous nuclear ribonucleoproteins (hnRNPs), which tend to inhibit (i.e., silence) splicing (Geuens, Bouhy, and Timmerman 2016; Blencowe 2006), and serine and arginine rich proteins (SRs), which tend to enhance splicing (Ule and Blencowe 2019), although the regulatory activity of SFs can be highly dependent on context and position of binding within the pre-mRNA (Marasco and Kornblihtt 2022). Many annotated SF functions have been produced by hypothesis-driven studies, and, more recently, systematic efforts, such as through the Encyclopedia of DNA Elements (ENCODE) project have mapped the functional and biophysical networks of 80 SFs in two human cell lines (Van Nostrand et al. 2020).

4.3.2 The emerging role of SFs in mediating pluripotency and differentiation

Just as with TFs, SFs can also act as master regulators of cell fate decisions during stem cell differentiation acting through SRNs (Wright and Ciosk 2013; Jangi and Sharp 2014). For example, *SRSF2* promotes the AS of exon 9 of the *NUMB* gene, creating an isoform of NUMB that specifically governs *NOTCH* (Y. Li et al. 2021), and thus inducing endothelial cell progenitor cell specification (Y. Li et al. 2021). Muscleblind-like splicing factor (*MBNL1*) and RNA Binding Fox-2 (*RBFOX2*) alternatively splice targets that drive iPSCs to mesoderm transitions (Venables et al. 2013). RBPs *ESP1*, *ESPR2*, *RBFOX2* and *QKI* coordinate an isoform switch that promotes tissue remodeling from a mesenchymal to epithelial state for kidney development (Wineberg et al. 2022). And, RNA binding motif protein 24 (*Rbm24*) drives cardiac differentiation programs in human and mouse by AS of genes related to cytoskeletal proteins and ATPase components that promote cardiac development (Zhang et al. 2016). Splicing regulator *QKI* is essential to establish splicing networks that control contractile and structural genes in the heart (Montañés-Agudo et al. 2023). The repression of PTBP1, another splicing regulator, can convert cardiac fibroblasts into cardiomyocytes or fibroblasts into neuronal cells, indicating its critical role in cell differentiation (Keppetipola et al. 2012; Xue et al. 2013). *PTBP1* influences transcriptional networks by regulating transcription factor *PBX1* during neuronal differentiation (Linares et al. 2015). *RBFOX2* is necessary for establishing splicing regulatory networks required for heart and neuronal development (Gehman et al. 2012). Overall, these examples highlight how SFs can serve as master regulators of cellular fate, similar to TFs.

SF networks can be characterized by experimental methods that link RBPs to the target isoforms they regulate. There are several different approaches that include RNA Immunoprecipitation (RIP)(Gagliardi and Matarazzo 2016), and enhanced crosslinking and immunoprecipitation (eCLIP) (Van Nostrand et al. 2016). Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) (Danan, Manickavel, and Hafner 2016). One of the most commonly used approaches is CLIP-seq, which links SF to their binding sites within their respective RNA targets transcriptome-wide is crosslinking and immunoprecipitation followed by RNA-sequencing, or CLIP-seq (Ascano et al. 2012; Hafner et al. 2021). Here, RBPs are subjected to UV irradiation, which cross-links RBPs to the RNA at the site to which they bind. The RBPs, which remain covalently bound to the RNA target, are enriched as an RNA-protein complex, and the population of bound RNA is deeply sequenced and subsequently aligned to the genome to depict locations of the bound RBPs (Hafner et al. 2021), revealing the positions across pre-mRNA and mRNA wherein the RBPs are bound, and presumably active, in a given condition or sample. Another method maps functional relationships between SFs and their targets by experimentally modulating RBP concentrations, such as through siRNA knockdown or overexpression plasmids, followed by measurements of splicing using RNA-seq (Van Nostrand et al. 2020).

4.4 Splice factors as regulators of the regulators - TF and SF isoforms

Among the many types of genes that splicing can target, the most marked effects are likely through alternative splicing of potent regulators of stem cell fate. Alternative splicing of such regulators could generate isoforms of the same gene with variable activities - from loss of function to gain in new functions (Castaldi et al. 2022). In other words, splicing can have outsize effects on stem cell fate by “regulating the regulators”.

4.4.1 Splicing influences GRNs by producing TF isoforms that differentially regulate cell fate

A continuum of functional relationships between TF isoforms of the same gene can occur, from isoforms with attenuated, opposite, or tandem functions. The levels and relative stoichiometries of such TF isoforms are directly influenced by AS, and thus AS can influence gene regulatory networks by modifying TF activities through splicing (López 1995; Niwa 2018).

AS can modulate TF functions by production of a sub-functional isoform. For example, AS modulates the activities of *OCT4*. *OCT4* produces at least three major isoforms (*OCT4A*, *OCT4B* and *OCT4B1*)(X. Wang and Dai 2010; Atlasi et al. 2008). *OCT4A* is responsible for establishing pluripotency. In contrast, an isoform switch to *OCT4B*, a sub-functional form of *OCT4*, results in inhibition of stem cell self-renewal, but may be involved in responses to cell stress (X. Wang and Dai 2010). The differences in isoform function may be due to differential inclusion of localization signals in which a nuclear localization signal in *OCT4A* is absent in *OCT4B*, reducing its nuclear residence and thus transcriptional activity (Cheong and Lufkin 2011).

AS can modulate TF function by producing an isoform with opposite function. For example, splicing can alter the specificity of the forkhead transcription factor (*FOXP1*) where one isoform promotes pluripotency by directly stimulating expression of the Yamanaka factors, while another isoform is predominantly expressed in differentiated cells and represses pluripotency genes (Gabut et al. 2011).

And, lastly, AS can modulate TF function in a way in which there is a division of labor between multiple TF isoforms of the same gene. An example can be found in *SALL4*, a member of the spat-like gene family (Tatetsu et al. 2016). *SALL4* interacts with both *OCT4* and *NANOG* to regulate pluripotency networks. *SALL4* produces two isoforms, *SALL4A* and *SALL4B*, which collaborate to maintain pluripotency networks (Rao et al. 2010). *SALL4A* represses genes associated with differentiation while *SALL4B* promotes pluripotent gene expression (Chepelev and Chen 2013). Interestingly, expression of *SALL4B* alone is not sufficient to promote the pluripotent state (Rao et al. 2010).

Analogously to how SFs can modulate activity of TFs, in a similar vein, SFs themselves may also act as upon themselves to modulate their own splicing (Jangi and Sharp 2014). Indeed, a pervasive mechanism of SF regulation is a negative feedback loop in which a SF binds to its own pre-mRNA which leads to nonsense-mediated decay products. While further studies work to elucidate these mechanisms as it relates to differentiation, negative regulation has clearly been demonstrated to affect 10-30% of mammalian genes (Jangi and Sharp 2014), and may play an important role in stem cell SRNs. GRNs and SRNs likely work in concert within stem cells to regulate pathways of pluripotency or differentiation (**Figure 4.1**).

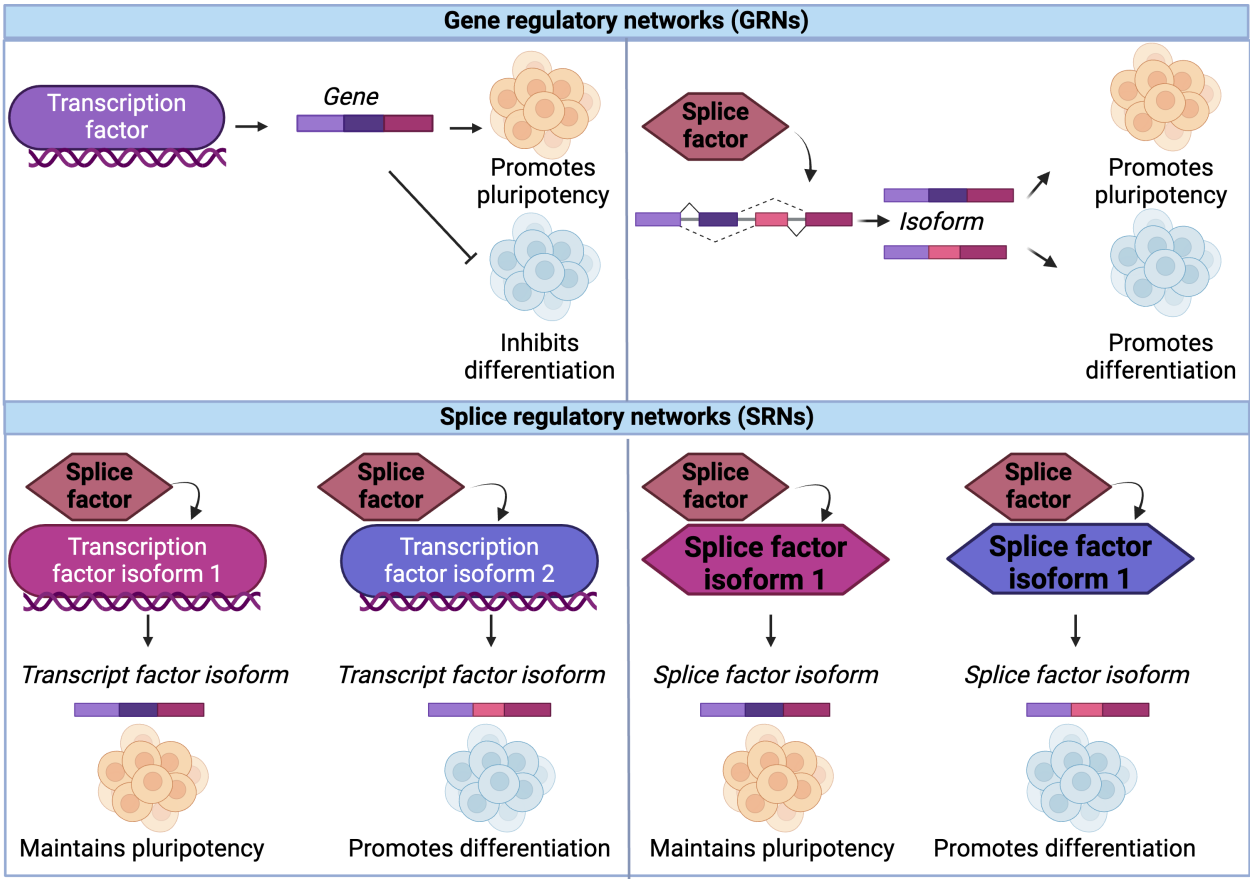


Figure 4.1 Depiction of the complex interplay between gene and splice regulatory networks
Gene regulatory networks (GRNs) are composed of transcription factors (TFs) and their target genes, controlling the expression of genes to determine pluripotency and differentiation. GRNs can also be shaped by alternative splicing (AS), resulting in the production of different isoforms with varying functions in stem cell fate. Splice regulatory networks (SRNs), consisting of splice factors (SFs) and their target splice sites, can have a significant impact on GRNs by producing different TF isoforms with different functions in stem cell phenotypes. SRNs can also regulate other SFs, leading to further control of the balance between pluripotency and differentiation.

4.5 Characterization of the transcriptome at isoform-resolution

GRN and SRN programs can influence transcriptome expression. For characterization of the stem cell transcriptome, early efforts employed mid-throughput real-time quantitative PCR (RT-qPCR)("Real-Time QRT-PCR" n.d.) assays to quantify pre-selected isoforms panels (Atlasi et al. 2008). More recently, short-read RNA-sequencing (SR RNA-seq) has enabled facile characterization of thousands of annotated and novel splice junctions and exon expression associated with stem cell phenotypes (Hardwick et al. 2019; Mortazavi et al. 2008; Conesa et al. 2016). However, the short length of RNA-seq reads limit observation of the entire unambiguous full-length isoform (Steijger et al. 2013; Alexey I. Nesvizhskii 2014).

These limitations are addressed with long-read (LR) sequencing platforms, such as through Pacific Biosciences Inc. (PacBio) or Oxford Nanopore Technologies (ONT) (van Dijk et al.

2018; Eid et al. 2009; Rhoads and Au 2015; Mantere, Kersten, and Hoischen 2019). In the decade following their introduction, the throughput, affordability, and accuracy of LR sequencing was lower than SR, but with the steady evolution of LR sequencing systems in terms of the chemistries, instrumentation, and computational pipelines (e.g., PacBio's Revio (Eid et al. 2009), ONT's iSeq100 (Jain et al. 2016)), accurate transcriptome sequences will likely become accessible at a large depth and breadth. With this higher depth, single cell LR (scLR) methods are being developed for both ONT and PacBio platforms. With ONT as a readout, Barcode identification from Long-reads for AnalyZing single-cell gene Expression (BLAZE) relies on barcodes from ONT long-reads to profile isoforms at single-cell resolution (Mantere, Kersten, and Hoischen 2019). PacBio has also reached the throughput needed for scLR through development of multiplexed arrays sequencing (MAS-ISO-seq), which uses a concatenation approach to ligate multipole cDNAs into large single molecules that is then sequenced (Al'Khafaji et al. 2023). To infer the functional interrelatedness of isoforms from scLR data, new analysis pipelines have been developed, such as *acorde*, which analyzes co-expression networks of correlated isoform abundances (Arzalluz-Luque et al. 2022).

4.5.1 Transcriptome and splicing regulation are an intertwined process

AS is co-transcriptional, and thus splicing and transcriptional biochemical processes work in concert within an epigenetic context to drive transcriptional outputs during differentiation (Tilgner et al. 2012; Kosti, Radivojac, and Mandel-Gutfreund 2012; Han et al. 2017). Chromatin state has a large influence on splicing (Kosti, Radivojac, and Mandel-Gutfreund 2012; Naftelberg et al. 2015). Two models have been proposed to describe how chromatin affects splicing outcomes: the kinetic and recruitment model. The kinetic model relates RNA polymerase II transcriptional speed with splice status. "Slow" Pol II increases the time with which an SF is exposed or could bind their cognate RNA binding sites, and thus promotes exon inclusion. The recruitment model focuses on the ability for components such as Pol II C-terminal tail or histone tails to mediate interaction-driven recruitment of SFs onto the nascent pre-mRNA which in turn affects AS (Luco et al. 2010; Agirre et al. 2021). Beyond the influence of chromatin and splicing, the biochemical relationships of splicing and transcription are surprisingly intertwined. TFs can bind to and regulate nascent RNA, influencing not only gene expression, but potentially RNA processing such as splicing (Han et al. 2017; Liang et al. 2022; Oksuz et al. 2022). SFs, on the other hand, can influence transcriptional regulation. For example, *RBM20*, through splicing, regulates genes necessary for heart development (Bertero et al. 2019).

Not only are the biochemical mechanisms of transcription and splicing intertwined, but components of the transcriptional (TFs, gene targets) and splicing (SFs, splice targets) networks feed into each other and are involved in cross-regulatory logic, in sometimes unexpected ways. The transcription factor *OCT4* upregulates the splice factor *SFRS2*, which regulates the splicing of methyl-CpG-binding protein, *MBD2*, whose isoforms play opposing roles in reprogramming to pluripotency (Lu et al. 2014). Upregulation of *SFRS2* increases levels of an isoform of *MBD2*

(*MBD2c*), which binds to the promoter of *OCT4* to reinforce the pluripotency core network. Interestingly, loss of either *OCT4* or *SFRS2* activity leads to product of the other isoform of *MBD2* (*MBD2a*), which also binds to the promoter of *OCT4*, but has a different C-terminal domain that silences *OCT4* expression by recruiting the Nucleosome Remodeling and Deacetylation complex (NuRD) (Lu et al. 2014). *SRSF2* does not just regulate *MBD2*, but can also change the activity of the transcription factor *FOXP1*, whose isoforms are nearly opposite in their induced stem cell phenotypes; one isoform of *FOXP1* activate genes that promote pluripotency and the other isoform promotes differentiation, a splicing switch that involves exonic changes to the *FOXP1* DNA binding domain (Gabut et al. 2011). Notably, TFs with zinc finger domains can also bind RNA. Zinc finger domains allow binding to both DNA and RNA. It has been shown that transcription factor *GATA4* can bind RNA and regulate alternative splicing networks in the heart (Zhu et al. 2022).

4.6 The functional output of the transcriptome - the stem cell proteome

In large part, the functional effect of the transcriptome manifests through the proteome, making identification of protein expression within the cell equally important. Indeed, transcript and protein abundances are not always highly correlated, with this relationship being affected by several factors including co- and post-translational layers of regulation (Aydin et al. 2022; Y. Liu, Beyer, and Aebersold 2016). Given the importance of splicing networks in stem cell differentiation, approaches to directly measure their functional outputs, or protein isoforms, in stem cells are critical (Kornblihtt et al. 2013; Tress, Abascal, and Valencia 2017a).

Mass-spectrometry (MS)-based proteomics is a powerful technique for comprehensive general proteome characterization of stem cell states (Lindoso et al. 2019; Gundry, Burridge, and Boheler 2011; J. Wang et al. 2008). MS has been used to track dynamic changes of proteins over stem cell differentiation, which may not correlate with transcriptional changes. An early work on transcript and protein levels in hESCs demonstrated that up to 50% of changes in protein expression do not have corresponding transcript changes, although some of this non-correlation is attributable to technical variability of the first-generation MS instruments (van Hoof, Krijgsveld, and Mummery 2012). The value of proteomics for discovering important stem cell factors continues to be demonstrated. In a study of young and adult mouse HSCs, a module of proteins were specifically expressed in young mouse HSCs, uncorrelated to the transcript levels (Van Hoof et al. 2008; Zaro et al. 2020). Multiplexed isobaric labeling, now a standard approach, allows for measuring protein expression along many more differentiation timepoints (Sabatier et al. 2021). Newer thermal profiling approaches coupled with MS can even assay intrinsic protein stability, and regulated destabilization during early differentiation, such as the ribosomal machinery that exhibits higher stability during differentiation (Sabatier et al. 2021).

Though general protein content from genes is readily measured, an ongoing challenge remains the characterization of the final output of gene and splicing regulatory networks: the proteome at isoform resolution. MS characterization of protein isoforms have been applied to

histone isoforms and isoforms in mesenchymal stem cells (Phanstiel et al. 2008; She et al. 2012). Technical challenges of isoform detection remain, though. In bottom-up MS, proteins are proteolytically digested into short peptides—few peptides uniquely map to an isoform (Blakeley et al. 2010), and such peptides are under sampled due to technical issues (MS ionization, charge, etc. (Marasco and Kornblihtt 2022)). New proteogenomic approaches (Alexey I. Nesvizhskii 2014; Parrotta et al. 2019; Miller et al. 2022; Sheynkman et al. 2016) can enhance protein detection accuracy and coverage by leveraging matched long-read RNA-seq data to generate a sample specific database of protein isoforms used for MS searching (X. Wang et al. 2012), which can provide direct protein level evidence of stable isoform expression (Mehlferber et al. 2022).

4.6.1 Emerging scalable strategies to causally link isoforms to stem cell phenotypes

Given the critical nature of SRNs in driving cell fates, determining the functional role of their downstream protein isoform products are critical, such as through experimentally testing the effect of knocking out or overexpressing an individual isoform on stem cell phenotypes. Isoforms can be "knocked down" using short interfering RNAs (siRNA) that are designed against regions specific for the target isoform mRNA (Endoh and Ohtsuki 2009; Dana et al. 2017). This approach has been used to modulate splicing patterns of isoforms in cancer cells to functionally decouple gene expression patterns from the individual role of isoforms in driving cell phenotype (Prinos et al. 2011).

To further modulate isoform expression, morpholino oligonucleotides can be used, which are RNA sequences designed to be complementary to the target sequences of RNA. Upon morpholino binding, spliceosome assembly or translation is inhibited through steric hindrance (Corey and Abrams 2001; Moulton 2007). And, the Type II CRISPR-based system Cas13 demonstrates the ability to knock-down RNA isoforms with high specificity and efficiency, with potential to design gRNAs against isoform-specific regions (Abudayyeh et al. 2017; Cox et al. 2017). The CRISPR-Cas9 system has been engineered to modulate expression of individual exons through the paired guide RNAs for alternative exon removal (pgFARM) to enable functional testing of individual exons that may be part of AS pathways (Thomas et al. 2020). Further application of the CRISPR system has been extended to the CRISPR (Artificial Splicing Factors (CASFx) system which provides the ability to induce AS events onto target regions, to mimic and functionally characterize specific splice isoforms (Du et al. 2020) .

For similar experimental goals, isoforms can be over-expressed to understand the contribution an isoform has in driving cell fate. Currently, large-scale overexpression screens have been applied to TFs by using a strong promoter to profile individual TF isoforms driving stem cell differentiation (Aragon and Hirschi 2022). Creation of "ORFeome" libraries have been created to functionally describe and characterize human isoforms (Yang et al. 2016). However, large-scale overexpression screens to functionally interrogate the role of isoforms in driving stem cell differentiation have yet to be performed, but similar methods could be applied to elucidate the greater role of isoforms in driving stem cell fate. These screens could complement massively

parallel reporter assays for screening candidate functional exonic or cryptic splice variants associated with stem cell mediated traits (Rhine et al. 2022; Soemedi et al. 2017). A range of splicing effects could be evaluated including, intron retention, exon exclusion, 5’ or 3’ UTR usage. More recent high-content cell imaging assays could be combined with these molecular functional screens (Veschini et al. 2021). By comparing different molecular splicing effects with gene expression and protein abundances and linking these to stem cell phenotypes, we expect these assays to dissect the complex architecture for a range of diseases such as cancers, cardiovascular disease, and autoimmune diseases.

4.7 Unresolved questions in the field and future directions

Stem cells are powerful model systems that mirror the processes of human development. Methods to date have focused on linking transcript expression signatures to stem cell phenotype, but emerging methods that combine multiple facets of regulation, such as transcriptional and splicing networks during stem cell differentiation should capture important regulatory programs previously missed. The interrelatedness of transcriptional and splicing networks, during biochemical regulation, as well as between their network components, during stem cell differentiation necessitates a multi-faceted approach to understand.

Chapter 5 Development of RNA-sequencing technologies to enhance characterization of the transcriptome

The content of this chapter is adapted from:

Transcript detection and quantification using Kinnex Full-length RNA Sequencing Data

David Wissel, University of Zurich, Joint work with **Madison M. Mehlferber** with joint supervision by Gloria Sheynkman (UVA) and Mark D. Robinson (University of Zurich)
https://programs.pacb.com/l/1652/2024-03-08/44gtlx

And

Kinnex full-length RNA kit for isoform sequencing

https://www.pacb.com/wp-content/uploads/Application-note-Kinnex-full-length-RNA-kit-for-isoform-sequencing.pdf

5.1 Evolution of next-generation sequencing platforms

In this chapter I will discuss the processing and analysis of PacBio long-read RNA-sequencing data and the recent advancements in technology within PacBio. I will also discuss how such advances in technology can further support efforts to enhance characterization of isoforms.

5.2 Long-read RNA sequencing technologies enhance resolution of the genome

Capturing high resolution information about isoform landscapes is critical for understanding their roles within tissues and cell states. As previously mentioned, isoforms harbor small stretches of unique sequences distinguishing them from their isoform counterparts, therefore utilizing technology that provides resolution of these small regions at high-confidence is imperative.

As discussed previously, the short-fragments provided by short-read RNA sequencing platforms are not ideal for all biological studies such as those for distinguishing isoforms. However, the third-generation of PacBio long-read RNA sequencing technologies offers support for isoform centric resolution. The technology itself is the key to supporting such research.

During PacBio long-read RNA-sequencing, after the cDNA molecules from the sample of interest receive SMRTbell adaptors forming circular DNA molecules, they enter the SMRT Cells receiving polymerase which initiates the sequencing reaction (refer to **Chapter 1 Figure 1.4B** and **Chapter 1.2.5** for a detailed description of PacBio SMRT sequencing). The polymerase

revolves around the formed circular DNA molecule, and during each revolution (10 on average) captures the sequence of the nucleotide. The raw reads are aligned to determine the circular consensus read (CCS) which is the average nucleotide collected during sequencing of the circular DNA molecule (refer to **Figure 5.1** for a visual depiction of CCS). These CCS reads with a high base-calling accuracy score (Q > 99%) become high-fidelity reads (HiFi) (Hon et al. 2020). The ability to combine several observations generated from multiple polymerase passes around the circular DNA molecule provides confidence in the true sequence of the transcript sequenced, resulting in highly-accurate reads.

The total number of reads generated from the Iso-Seq Sequel IIe platform generally produces on average about 3 million HiFi reads (Al'Khafaji et al. 2023). These reads can then be mapped to the genome, guided by reference transcriptomes, to generate the set of distinct transcripts and genes and their respective abundances. Long read alignments can be further analyzed with downstream open access tools, discussed in section **5.2.3**.

While the read depth (or total number of reads) of long-read RNA sequencing runs are generally lower compared to short-read RNA sequencing, it is important to note that third generation sequencing technologies generate full-length transcript sequences spanning the 5' to 3' end of a transcript, eliminating the need for probabilistic transcript assembly as with short-read RNA sequencing approaches. Thus, full-length transcripts allow for highly accurate identification of transcripts, limiting ambiguity of sequence identity, including the discovery and annotation of novel full-length isoforms.

However, on the PacBio Sequel IIe platform, it has been observed that for optimal base calling quality to be achieved, larger library sizes are needed (15 – 20kbp) to allow for optimal sequencing, which can be difficult to achieve with highly precious or limited samples (Al'Khafaji et al. 2023). With the transcriptome serving as a proxy for diagnostics of gene expression within a sample, refined resolution of the genome for sensitive and biomedically relevant samples can greatly support advancements in health. Therefore, efforts to increase throughput and resolution are still ongoing. Increasing throughput within the PacBio platform means the generation of more HiFi reads, catapulting confidence in the identities of the transcripts identified and at a greater scale.

5.2.1 Development of MAS-Iso-Seq to increase RNA-sequencing throughput

The most recent advancement to improve throughput of PacBio sequencing is the development is of MAS-Iso-Seq, created by Dr. Aziz Al'Khafaji et. al in 2023 which was later adapted by PacBio in 2023 and coined Kinnex.

In the Kinnex method, cDNA libraries are separated into individual PCR reactions to increase yield, followed by the addition of unique barcode adaptors to each library. After amplification, the barcodes are used to ligate together samples from each library, then hybridized to form a concatemer of several cDNA molecules (up to 8 currently). In the final step, SMRT

bell adaptors are added to the terminal ends of the concatenated cDNAs to coerce the formation of the circular DNA molecule needed for SMRT sequencing and obtaining HiFi reads. The concatenation step is what allows for the increase in throughput, enabling the sequencing of up to 8 cDNAs within one circularized molecule rather than one cDNA sequenced per molecule, as in the original Iso-Seq method (**Figure 5.1**).

The sequencing reaction following the Kinnex protocol, occurs as described, collecting circular consensus reads (or HiFi reads). Rather than revolutions of about 10x per circular DNA molecule, the Kinnex approach utilizes a higher fidelity polymerase performing on average 16 revolutions per circular DNA molecule further enhancing throughput to increase the number of reads generated. Overall, this approach enables a 3x increase in the number of HiFi reads collected (**Figure 5.1**). It is important to note that within that HiFi read, there are 8 transcripts being simultaneously sequenced, meaning that there is an 8x increase in the number of transcripts being sequenced. When multiplying this over the entirety of the SMRT Cell, this yields a 15x increase in the number of total reads generated via Kinnex. The resulting reads from Kinnex can be bioinformatically de-constructed (de-concatenated) to identify the individual transcripts comprising the amplicon molecules.

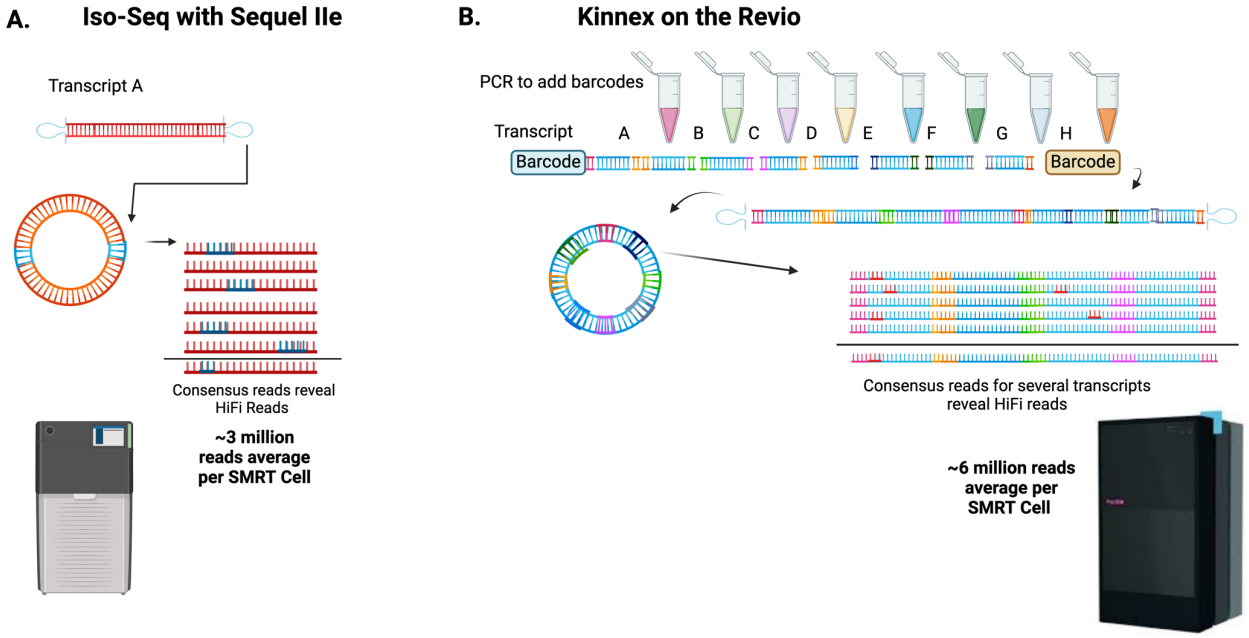


Figure 5.1 Evolution of PacBio HiFi long-read RNA-sequencing SMRTbell library preparation

In the Iso-Seq method (A), one cDNA molecule is transformed into one circular DNA molecular that will undergo single-molecule real-time sequencing (SMRT sequencing). The resulting reads are used as the input for the Iso-Seq analysis platform to define a high-confidence (HiFi) read. In the Kinnex platform (B), rather than 1 read comprising 1 circular DNA, multiple libraries of cDNA are separated into individual PCR reactions to append barcode adaptors to the ends of the individual cDNA libraries. After amplification of the libraries, the barcodes are ligated together forming a concatemer of several cDNA molecules. Overall, the Kinnex platform results in a 15x increase in the number of reads generated from

previous technology (based on CCS reads). Collected sub-reads are bioinformatically processed and separated by their barcodes to obtain the set of distinct transcripts.

The Kinnex workflow was developed on the Sequel IIe resulting in 15-20 million reads. The introduction of the PacBio Revio sequencer in 2023 supported further enhancements in throughput and when combined with the Kinnex protocol generated 37- 40 million reads for downstream analysis. Improvements in throughput were achieved in multiple ways. Not only did the Kinnex workflow enable the parallel sequencing of multiple transcripts with the introduction of the concatenation technique through MAS-Iso-Seq, but technological considerations were included within the Revio that supported increased throughput as well. The Revio was accompanied by the introduction of a new SMRT Cell with 25 million ZMWs and the capability to sequence up to four SMRT Cells in parallel making 100 million ZMWs available for sequencing runs. The Revio sequencer also includes the accompaniment of the NVIDIA GPUs allowing for a 20-fold increase in computing power. Together, the combination of the Revio with the Kinnex protocol and enhanced ZMW chips has allowed for over an 8-fold increase in throughput from previous protocols. Thus, this technological advancement resulted in about 15x increase in the number of reads generated from previous technology (Iso-Seq and Sequel IIe, **Figure 5.1)**.

With increased throughput comes enhanced sensitivity and additional coverage of the transcriptome. Specifically, increases in the number of reads enhance our confidence in the sequence identity.

5.2.2 Bioinformatic tools for analysis of PacBio-derived data

Thus far I have only discussed the sequencing methods and platforms for collecting long-read RNA-sequencing data, however data analysis is a key element. Here I will discuss the suite of software for long-read RNA-sequencing data to support isoform discovery. However, the PacBio Github (https://github.com/PacificBiosciences/pbbioconda) provides a comprehensive list of packages, beyond those described in this thesis, to support applications of PacBio RNA-sequencing

The term *Iso-Seq* not only applies to the method of sample preparation for collecting long-read RNA-sequencing data but also the bioinformatics processing steps needed to generate final transcript expression files. The *Iso-Seq* analysis pipeline consists of five major steps (**Figure 5.2**) (Gordon et al. 2015). The first step is *Collapse,* which combines the multiple HiFi reads to generate consensus reads. *Demultiplexing* follows to remove the barcodes added during the sequencing reaction and filters out suboptimal reads utilizing tools, such as *Lima*. The *Refine* process follows offering additional filtering opportunities removing residual sequences from barcodes or filtering transcripts that are potential artifacts formed during the SMRT bell adaptor step. The next step is *Cluster*, which takes the cleaned transcript sequences and generates an alignment to group similar transcripts together based on sequence similarity. *Mapping and*

Collapsing is the final step, aligning clustered reads to the reference genome utilizing packages, such as *pbmm2*, to determine the genes and transcripts represented within the data.

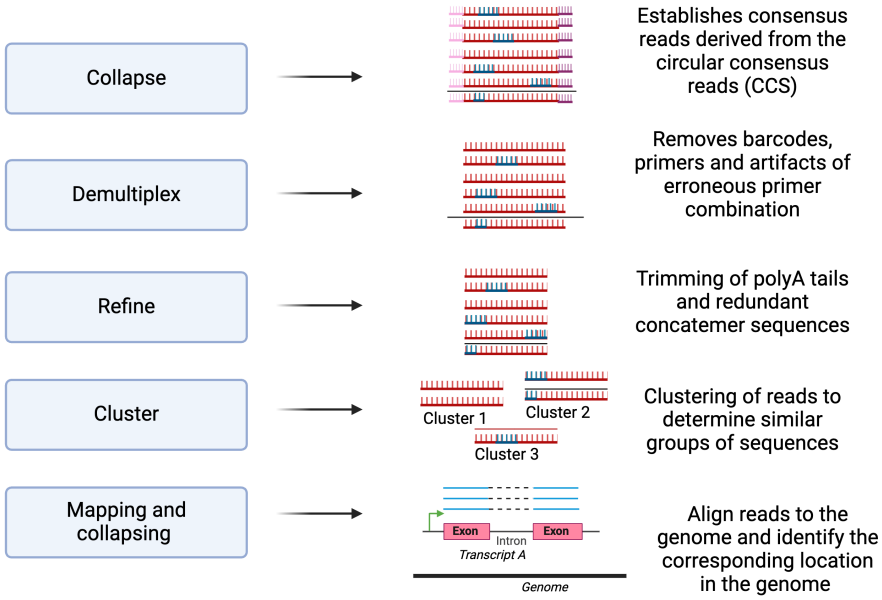


Figure 5.2 Overview of PacBio Iso-Seq workflow for processing of PacBio derived sequencing data
Overview of the Iso-Seq bioinformatic processing workflow and visual depiction of how such steps process the collected transcripts.

Once data is processed, downstream analysis options follow with a variety of open-source computational software. One common analysis for long-read RNA sequencing bioinformatic workflows includes the Structural and Quality Annotation of Novel Transcript Isoforms (SQANTI) (Tardaguila et al. 2017) which is a an open-source transcript classification tool to determine from a sample of long-read collected data the portion of transcripts that match reference annotation sequences versus transcripts categorized as novel. The tool "pigeon" is an implementation of SQANTI to PacBio specifications (https://isoseq.how/classification/workflow). The number of transcripts that represent true positives is an ongoing topic of discussion. However, projects such as the Long-Read Genome Annotation Assessment Project (LRGASP) have assessed various long-read RNA sequencing workflows. Such research efforts have suggested the need for orthogonal experimental approaches to validate the novel isoforms discovered during such analysis (Pardo-Palacios et al. 2024).

Another component of the LRGASP project included benchmarking of transcriptome analysis tools available for long-read RNA-sequencing data. Sub-aims of this effort included comparing tools for transcript discovery and quantification, recognizing accuracy in these competencies affects the datasets used for downstream data interpretation. Focusing on the top-performing tools as revealed through the study (Pardo-Palacios et al. 2024), IsoQuant and

Bambu emerged as top performing tools for transcript quantification. Each tool focuses on handling unique aspects of long-read RNA-seq data. IsoQuant (https://github.com/ablab/IsoQuant) is a Python package focused on accurately providing isoform discovery and quantification to support projects focused on alternative splicing by providing error correction features to improve accuracy. Bambu (https://github.com/GoekeLab/bambu) is an R package providing transcript quantification and novel transcript discovery. Bambu incorporates a novel machine-learning algorithm to learn features of bona fide transcripts and establish read-classes based on transcript similarity (Chen et al. 2023). A novel discovery threshold is then applied to discern between transcripts that are products of sequencing artifacts or actual novel transcripts (**Figure 5.3**).

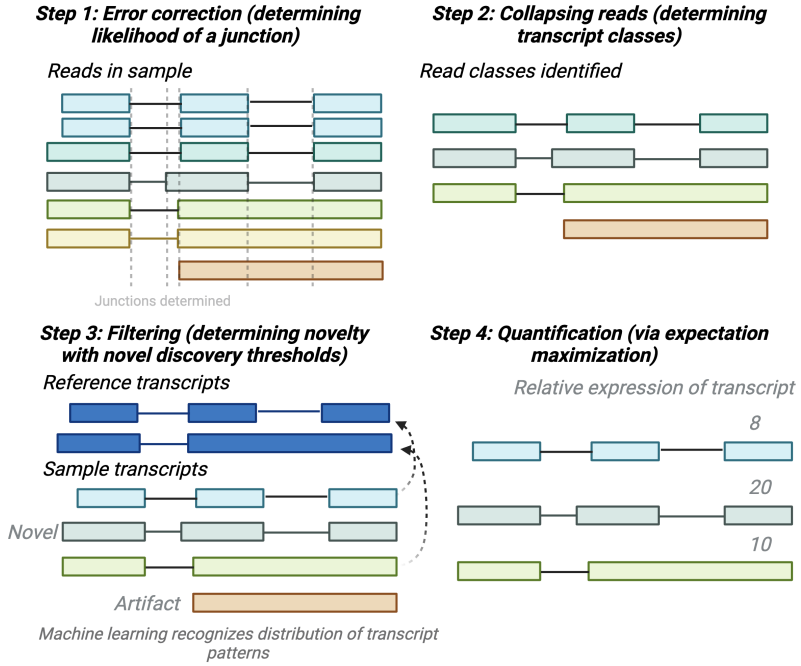


Figure 5.3 Overview of Bambu processing for quantification
Resulting data frames after executing Bambu are available and amenable for downstream applications to support analysis efforts of various applications.

5.3 Transcript quantification and isoform discovery with the new Kinnex platform

We obtained early access to the Kinnex platform to collect deep coverage long-read RNA-sequencing data on the Revio. We wanted to understand how Kinnex could support RNA-sequencing applications such as isoform discovery and measurement of differential transcript expression. Given the novelty of the Kinnex platform, we wanted to understand if this new technology enabled increased isoform resolution to support downstream isoform-related analysis.

To investigate the performance of Kinnex, I collected long-read RNA-sequencing data from induced human pluripotent stem cells (iPSCs, WTC11s) across multiple timepoints as they differentiated to become primordial endothelial cells, to characterize temporal isoform dynamics.

5.3.1 Sample collection for Kinnex RNA-sequencing

To obtain the dataset utilized for addressing these questions, WTC11 stem cells were subjected to a primordial endothelial cell differentiation protocol (Nelson et al. 2021) transitioning from pluripotency to then form a primitive streak, a mesodermal population, and finally primordial endothelial cells (ECs), over the course of 6 days. In order to capture changes in isoform expression across development, cells were collected at every day of differentiation with biological replicates for Day 0, 3 and 5. In order to assess the accuracy of PacBio Kinnex long-read RNA sequencing for isoform discovery, following RNA extraction, Spike-In RNA Variants (SIRVs, Lexogen (Paul et al. 2016)) were added as control transcripts with Day 0 biological replicates receiving E1 SIRVs, Day 5 E2 and all other Days receiving S4. Samples from each day were split into equal aliquots for parallel Kinnex long-read RNA-sequencing on the Revio and Illumina (Kapa RNA HyperPro Kits) short-read RNA-sequencing on the NovaSeq at 150 base pair (bp) paired end. Additional, details on stem cell culture methods and sample collection can be found in **Chapter 6.4 Methods**.

5.3.2 Data analysis

The dataset utilized for evaluating the effectiveness of Kinnex data for the study of isoform populations encompasses a 6-day differentiation time course. However, for the purposes of evaluation, the data discussed in this chapter will follow a 2-condition comparison between Day 0 (WTC11 cells) and Day 5 (primordial ECs) to align with the general format for many open-source analysis tools. Results of isoform dynamics during the entire primordial EC differentiation process are discussed in **Chapter 6**.

To compare sequencing coverage and read depth metrics, for all samples with Kinnex data, short-read RNA-sequencing data was collected in tandem for the same samples (**Figure 5.4**). To benchmark the accuracy of isoform discovery and differential expression, Lexogen SIRV Spike-In RNA Variants (SIRVs, Lexogen) served as control transcripts. SIRVs are engineered to map to non-human genes, allowing for separation of sample-derived transcripts verses control spike-in RNA data.

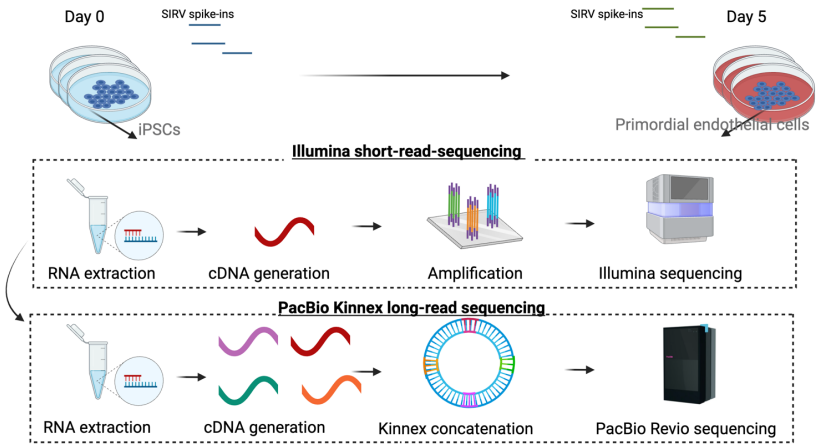


Figure 5.4 Experimental schematic of Kinnex data collection during iPSC to primordial EC differentiation
Samples were collected for benchmarking Kinnex long-read RNA-sequencing data with the Illumina short-read RNA-sequencing performed in parallel.

Overall, we observed on average a 2x fold increase in the number of HiFi reads generated between samples sequenced using the Non-Kinnex (Iso-Seq protocol) on the Sequel II/IIe as compared to Kinnex on the Revio (**Table 5.1)**. Additionally, we observed about a 3.5x increase in the number of S-reads generated with the Kinnex prepared samples when sequenced on the Revio compared to the Sequel IIe (**Table 5.1**).

Sample	Library	HiFi Reads	S-Reads
UHHR	Non-Kinnex-Sequel II/Iie	3,194,311	n/a
	Kinnex/Sequel II/Iie	2,720,033	20,453,854
	Kinnex/Revio	6,546,645	47,250,258
HG002	HG002	5,984,046	38,740,671
WTC11	Day0-rep1	6,920,750	54,110,504
	Day0-rep2	8,611,025	67,547,611
	Day0-rep3	8,124,744	63,251,235
	Day1-rep1	6,430,958	49,897,067
	Day2-rep1	7,353,759	58,217,895
	Day3-rep1	5,483,994	42,173,159
	Day3-rep2	6,687,580	52,317,384
	Day4-rep1	7,295,962	57,061,795
	Day5-rep1	6,645,009	51,741,094
	Day5-rep2	7,542,604	59,092,202
	Day5-rep3	6,358,300	49,466,302

Table 5.1 Number of reads generated via Kinnex on the Revio sequencer vs. previous long-read RNA-sequencing platforms
Comparison of collected reads (HiFi Reads) obtained via the Kinnex platform sequenced on the Revio and Iso-Seq Sequel IIe platforms for various cell lines and associated sub-reads (S-reads). S-reads and HiFi sequence numbers are calculated via the Iso-Seq workflow. Reference *Kinnex full-length RNA kit for isoform sequencing*.

Due to the mechanism by which the Kinnex protocol concatenates multiple cDNA's into one larger molecule to generate libraries, concerns might arise that there is a potential bias with shorter cDNA molecules being incorporated into the constructed cDNA molecule more efficiently. Despite the concatenation technique featured within the Kinnex protocol as described previously (**Figure 5.1**), we did not observe a skew in terms of the transcript lengths generated when comparing against a diverse range of samples (**Figure 5.5**).

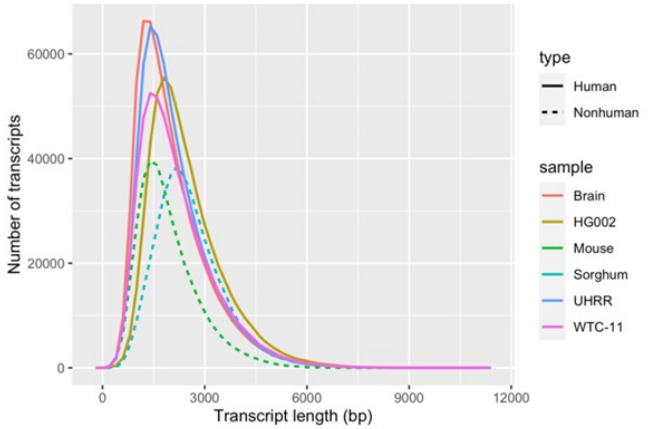


Figure 5.5 Distribution of transcript lengths is not affected by the Kinnex concatenation method (refer to *Kinnex full-length RNA kit for isoform sequencing*)
Transcript length distribution of various library samples created via the Kinnex protocol showing variations in length occur in a sample-specific manner but exhibit consistent size ranges. Figure made by Elizabeth Tseng (PacBio).

To measure transcript differences across the differentiation time course, we utilized Bambu (Chen et al. 2023) to quantify the sequencing data and derive count matrices to measure transcript differences across the differentiation time course (**Figure 5.6**). Bambu is ideally suited for the analysis of multi-sample PacBio data, as uniquely identifying accessions assigned to a transcript during sequencing vary between runs, limiting the ability to compare samples. Bambu enables construction of multi-sample quantification files.

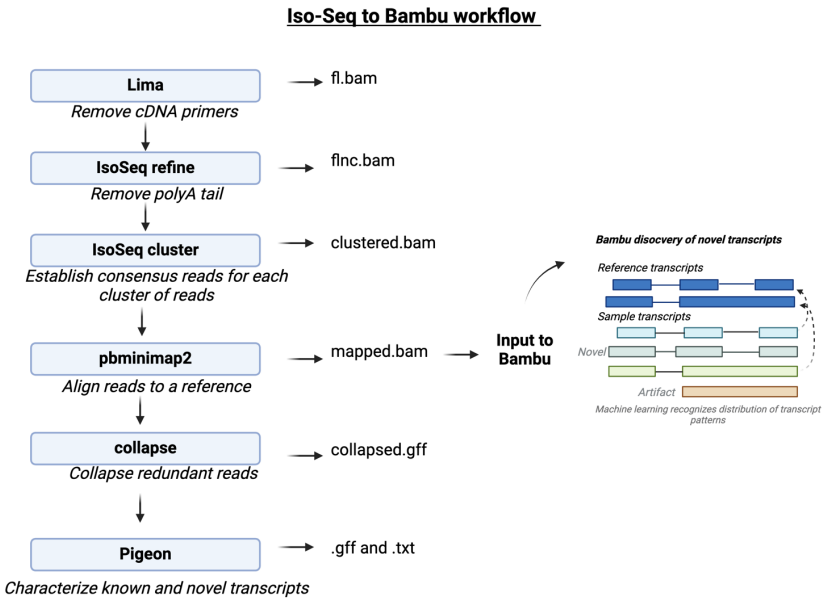


Figure 5.6 Workflow for utilizing Bambu to quantify and detect isoforms from Kinnex long-read RNA-sequencing data
Bambu works with the Iso-Seq Bioinformatic workflow, using as input the aligned bam files (*mapped.bam)* from pbminimap2, for quantification and novel discovery.

For a transcriptome-wide, global comparison of the results of Illumina versus Kinnex data, we determined the consistency of transcript expression within each platform by day. To assess this, we plotted the Pearson correlation between computed transcript abundances for Illumina and Kinnex (**Figure 5.7**). The correlation between Day 0 and Day 5 for Kinnex and Illumina are moderately high (0.75-0.8), indicating consistent expression patterns by day within each platform (intra-platform variability) (**Figure 5.7**). However, the correlations between Kinnex and Illumina are somewhat lower, suggesting that Kinnex captures transcript expression patterns comparable to Illumina, although with some variability. Such variability associated with intra-platform correlations has been previously observed (Pardo-Palacios et al. 2024).

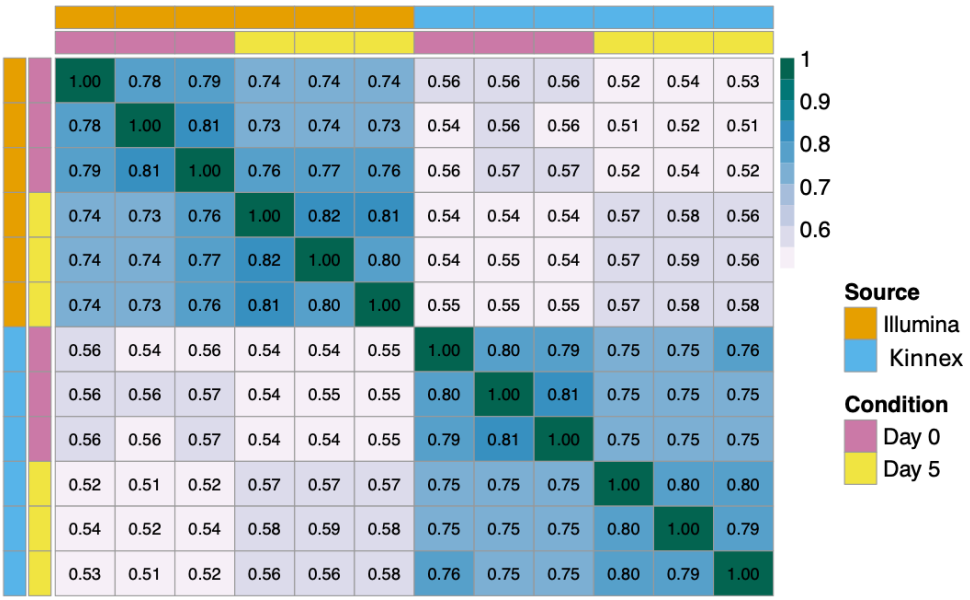


Figure 5.7 Kinnex quantification results shows competitive replicability to that of Illumina. Pearson correlation matrix representing transcript expression profiles between Day 0 and Day 5 faceted by Illumina and Kinnex technologies. Figure made by David Wissel.

For both the Kinnex and Illumina datasets, the number of reads returned was comparable. But, lengths of reads produced via Kinnex, on average are longer than those produced by Illumina, consistent with our expectations for long-read RNA-sequencing results (**Figure 5.8A**). This means that though Kinnex and Illumina return roughly the same number of reads, the absolute number of bases that are sequenced by Kinnex is on average 2-fold greater (**Figure 5.8B**). And lastly, we confirmed that Kinnex concatenation does not alter significantly transcript length when compared with Iso-Seq technology (**Figure 5.8C**).

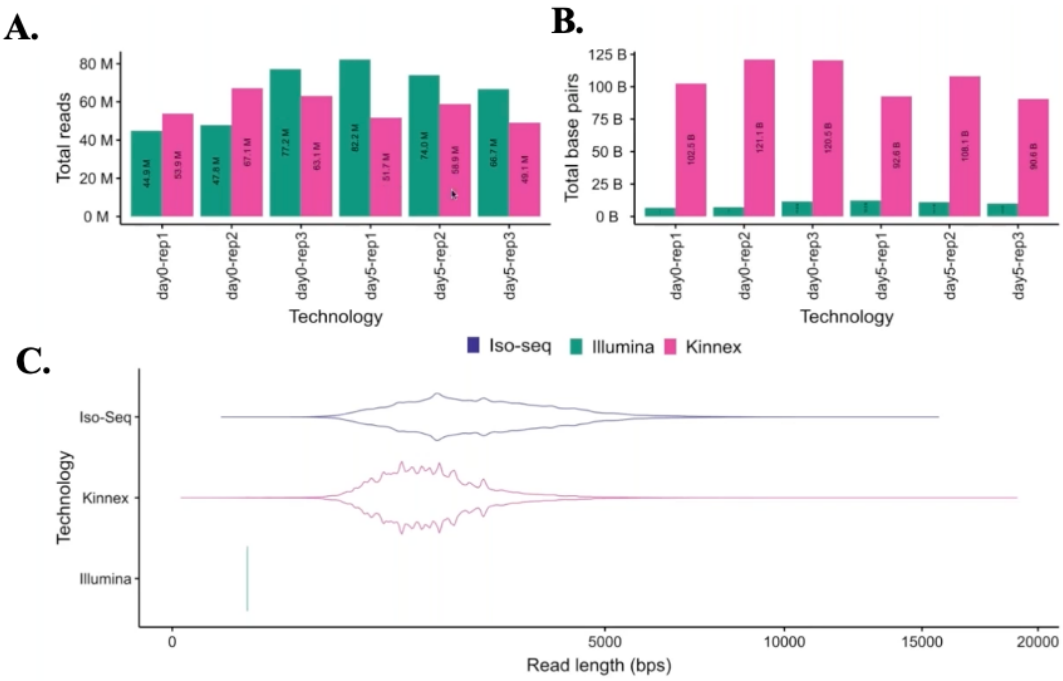


Figure 5.8 Kinnex provides more base pairs compared to Illumina due to its longer read-length
Bar plots showing the total number of reads (A) and base pairs (B) identified via Illumina and Kinnex platforms (C) Violin plots that display the distribution of read lengths for non-Kinnex Iso-Seq, and Illumina sequencing. Figure made by David Wissel.

The data we collected represents a dramatic transition between pluripotent and differentiated primordial ECs; therefore, I asked which genes and isoforms may be differentially expressed for Day 0 and Day 5 samples. We used the database PanglaoDB (Franzén, Gan, and Björkegren 2019), compiled from single-cell RNA-sequencing experiments from both human and mouse to provide a list of genes we expect to be differentially expressed between Day 0 and Day 5 samples. Overall, both Kinnex derived data and Illumina agreed in revealing a consistent set of genes up-regulated between the 2 conditions. For known EC markers, we found that overall genes with differential expression were concordant with expectations for gene expression programs associated with early EC phenotypes (**Figure 5.9**).

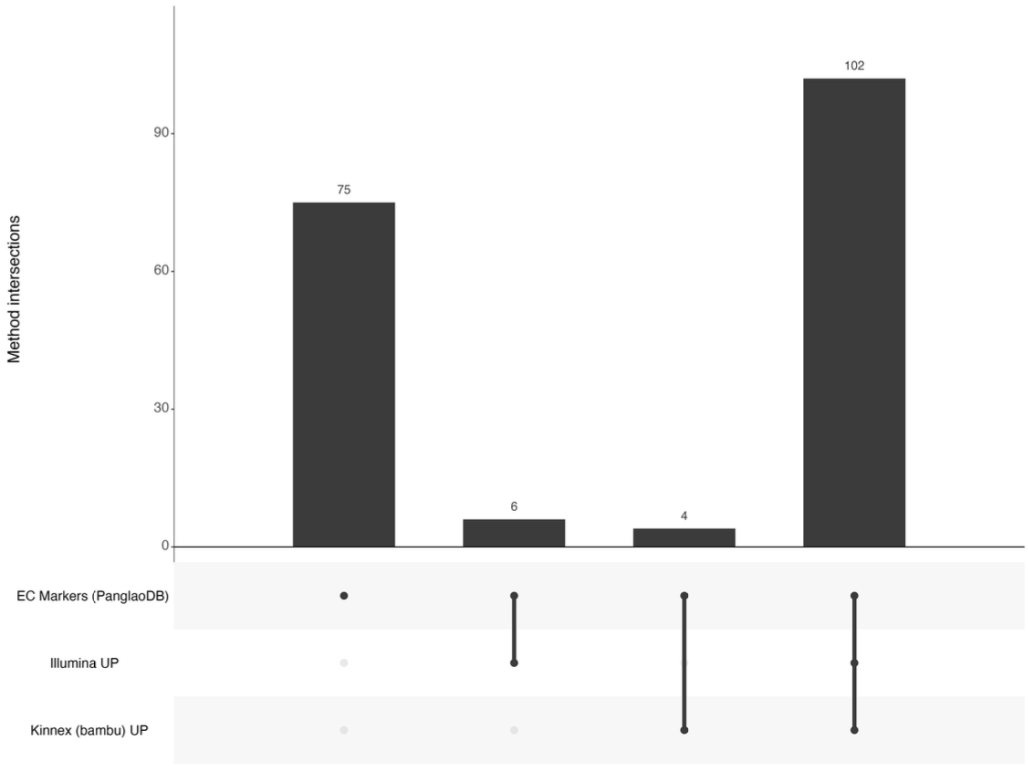


Figure 5.9 Differential gene expression of Illumina and Kinnex Day 0 and Day 5 show overlap between upregulated endothelial cell markers
Upset plot comparing the differentially expressed genes identified between Day 0 and Day 5 found on Kinnex and Illumina revealing a large proportion of recovered genes via both platforms. Figure made by David Wissel.

With these results, we were encouraged that the Kinnex platform was able to support efforts to comprehensively study isoform expression dynamics over the differentiation time course of primordial EC establishment described in **Chapter 6**.

Chapter 6 Deep coverage, high accuracy long-read RNA sequencing to characterize isoforms across early endothelial cell development

This chapter contains preliminary unpublished data:

Madison M. Mehlferber, David Wissel, Vasilii Pavelko, Elizabeth Nelson, Emily Watts-Whitehead, Erin D. Jeffery, Mark D. Robinson, Leon Sheynkman, Gloria M. Sheynkman

List of additional files:

- Appendix C: Supplemental Figures S1-S9

6.1 Introduction

Previously, the study of isoforms within this thesis has been assessed in a single cell line within HUVECs (see **Chapter 3**). As demonstrated in **Chapter 4**, isoforms contribute to the earliest establishment of cell fates, working to orchestrate discrete developmental transitions, with some splice networks governing division between mature and plastic cellular states (Gabut et al. 2011). Therefore, studying isoforms within a multi-sample, dynamic context could inform on the versatility achieved through isoform regulation within biological systems.

However, the extent of profiling isoforms across a temporal trajectory has been relatively narrow, limiting knowledge on the series of choices associated with cellular fate specification. It has been understood that isoforms modulate their expression during specific times of development to become defining features of specific developmental states (Mazin et al. 2021). But systematically tracking the changes in cellular decisions resulting at arrival of specific cell fates has not been widely documented.

Specific information on temporal isoform dynamics has been extremely limited in the vascular biology field, where comprehensive understanding of the factors underlying the development and establishment of early endothelial cell (EC) phenotypes is poorly characterized (Kelly and Hirschi 2009). Studies have shown that certain functions of ECs can be modulated through isoforms (Giampietro et al. 2015; Blanco and Bernabéu 2012; Park, Sorenson, and Sheibani 2015); however global characterization of isoforms involved in EC development has not been done in a comprehensive manner, nor at full-length isoform resolution.

For in-depth profiling of full-length isoforms in EC differentiation, we subjected human induced-pluripotent stem cells (WTC11s) to a five-day differentiation protocol to model the transition from pluripotency to primordial ECs, the precursors to mature EC populations and study isoforms that may be involved during this process. To delineate the factors involved in its regulation, we collected cells over six days of differentiation with biological replicates for Day 0, Day 3, and the final day, Day 5.

To construct an atlas of full-length isoforms associated with EC differentiation, we capitalized on increased depth of Kinnex long-read RNA-sequencing, described in **Chapter 5,** which enabled the collection of deep-coverage, full-length transcriptomics data for multiple samples. We identified isoforms with expression changes over the differentiation time course, including cases of isoform switching events. We highlight dynamically regulated isoforms previously implicated within vascular and EC pathways.

This work represents to our knowledge the first-comprehensive long-read RNA-sequencing atlas of differentiating ECs, at high splicing resolution. This work should serve as a resource for defining the genes and full-length isoforms potentially associated with developmental transitions.

6.2 Results

6.2.1 In vitro system to derive primordial endothelial cells from induced pluripotent stem cells (WTC11)

We surmised that using time course derived long-read RNA sequencing data could expose dynamic isoform expression patterns correlated with primordial EC developmental trajectory. Such information would form the basis of an atlas representing temporal expression profiles (**Figure 6.1A**). To construct the database of isoforms associated with this process, we designed a tractable system in which samples were collected every day during primordial EC development to analyze isoform expression (**Figure 6.1B**). We used a previously established protocol to generate primordial ECs from iPSCs in five days (Nelson et al. 2021). Utilizing small molecules, we coaxed cells from a pluripotent state (Day 0) to primitive streak (Day 1) with addition of glycogen synthase kinase (GSK3). Mesodermal lineages (Day 3) were created with addition of basic fibroblast growth factor (bFGF), and finally, a primordial EC phenotype (Days 4 and 5) was promoted via a combination of bone morphogenetic protein 4 (BMP4) and vascular endothelial growth factor A (VEGF-A) (**Figure 6.1B**). Additional details pertaining to cell culture can be found in **6.4 Methods**.

To confirm the progress of differentiation, we performed quantitative polymerase chain reaction (qPCR) analysis for gene markers associated with cell fate changes. We observed a decline in pluripotency genes (*SOX2*) while mesodermal and endothelial associated gene markers (*TBXT* (Brachyury protein), *HAND1*, *CDH5* and *VEGFR2*) increased in their expression across differentiation (Qiu et al. 2020) (**Figure 6.1C**). Given that samples expressed the expected genetic markers associated with their differentiation stage, we proceeded to use these samples for Kinnex sequencing to generate a long-read RNA-sequencing dataset.

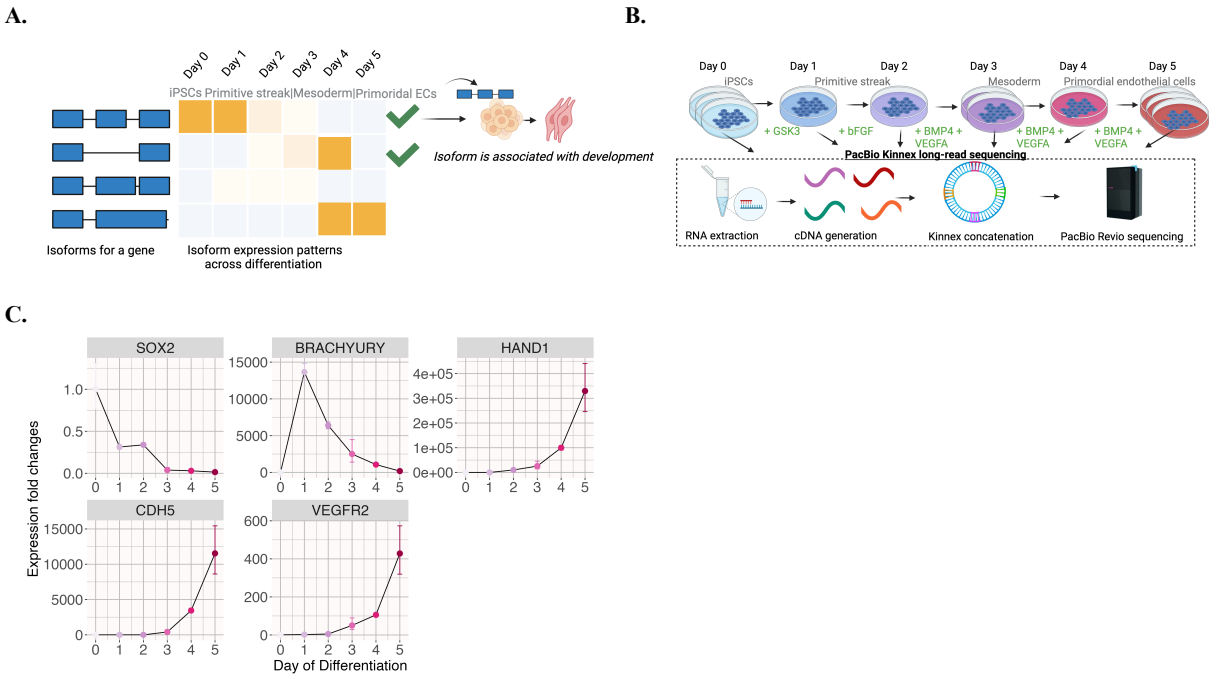


Figure 6.1 Use of an in vitro model of primordial endothelial cell differentiation from induced human pluripotent stem cells (WTC11)
A. Rationale for studying isoforms involved in differentiation to establish a temporal database of isoforms B. Experimental design to generate temporal isoform expression profiles of differentiating primordial endothelial cells C. qPCR validation of genetic markers associated with cell fate establishment with pluripotency (*SOX2*), early mesoderm (*TBXT*), mesoderm (*HAND1*), early endothelium (*CDH5*, *VEGFR2*)

6.2.2 Characterization of RNA-sequencing results

We utilized Bambu (Chen et al. 2023) to quantify transcript expression between the multiple samples to generate a comprehensive gene and transcript count matrix to track changes in transcript expression (described previously in **Chapter 5**). We processed Kinnex data for 11 samples (includes variable number of replicates of Day 0, 1, 2, 3, 4 and 5) and identified 60,000 genes and 254,000 isoforms. In this case, the term "gene" refers to a region in which there is a pile-up of full-length reads that is non-overlapping with other pile-up regions.

We wanted to ensure that Bambu quantification of the long-read RNA-sequencing PacBio data was accurately recapitulating expected gene marker expression. We performed an *in-silico* qPCR to focus on the genetic markers we utilized in the experimental qPCR to examine if gene expression from our constructed matrix matched similar trends. Overall, the Bambu derived gene matrix returned similar trends to those found experimentally (*Appendix C: Supplemental Figure S1*).

We calculated the Pearson correlation to measure transcript expression reproducibility across replicates and days. We found very high reproducibility among replicates from the same day (R^2~1) (**Figure 6.3A**). As differentiation progressed, we observed increased differences in

transcript expression profiles, as evidenced by the graduate decrease in correlation coefficients (**Figure 6.3A)**. Samples were exhibiting transcriptomic differences distinct from the prior day while becoming more similar to the subsequent day (**Figure 6.3A**). Since this analysis is based on a differentiation time course from bulk samples, and not single-cell resolved, not all cells are transitioning simultaneously, we expected to see variation in the observed correlation coefficients between days and samples.

We initially used Bambu as a quantification tool, but it also incorporates a novel discovery threshold to identify novel transcripts from long-read RNA-sequencing data. We identified 204 novel transcripts corresponding to 70 novel genes. These genes correspond to genomic regions without previous annotation within GENCODE. To provide a more detailed classification of the features leading to the transcript being classified as novel, we utilized the tool SQANTI3 (Pardo-Palacios et al. 2023) and further classified a total of 2,034 novel transcripts with 1,406 novel not in catalog (NNC) and 628 novel in catalog (NIC) corresponding to 1,187 genes. Overall, the known isoforms are in higher abundance than their novel isoform counterparts, but novel isoforms still are well represented within the data (**Figure 6.2B**).

Driven by the goal to profile the isoform landscape in early ECs, we focused on the frequency of genes (counts per million (CPM) >1) expressing different numbers of isoforms. We observed a large portion of genes were expressing multiple isoforms, with 9,227 genes expressing more than one isoform (**Figure 6.2C**). We also faceted genes expressing multiple isoforms by individual day and observed similar distribution patterns (*Appendix C: Supplemental Figure S2A*). Over the time course, however, we noted fluctuations in the number of genes that expressed multiple isoforms (*Appendix C: Supplemental Figure S2B*)

Next, we wanted to understand the types of splicing events (e.g., exon skipping) that were contributing to EC isoform diversity. We utilized the tool SUPPA (Trincado et al. 2018) to quantify splicing events from our Bambu-computed transcript expression. We observed a diversity of splicing patterns with a substantial proportion of genes expressing more than one splicing event type (**Figure 6.2D**). Skipped exons (SE) events were the most prevalent, followed by alternative 3; (A3) and 5’ (A5) splice site usage, while retained introns (RI) were the least common.

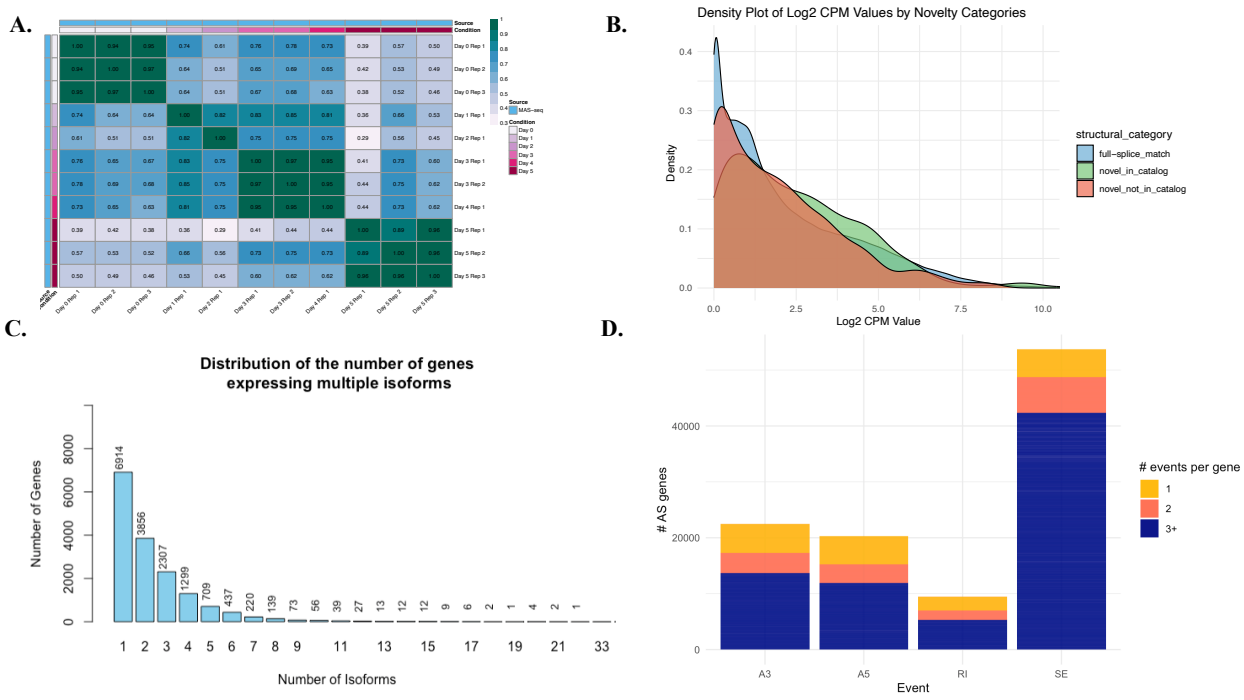


Figure 6.2 Characteristics of long-read sequencing
A. Pearson correlation plot comparing the relationship among replicates over the differentiation time course B. Distribution of the known and novel isoforms C. Distribution of the number of genes expressing multiple isoforms D. The breakdown of the number and type of splicing patterns exhibited (A3 = Alternative 3' splice site, A5 = Alternative 5' splice site, RI = Retained intron, SE = Skipped exon).

6.2.3 Profiling dynamic isoform expression patterns

Given the sheer number of isoforms profiled, we needed a systematic way to measure isoform changes that were the most significantly different over the time course. Gene-level expression analysis does not capture transcript-level changes that may be the result of alternative splicing (Marques-Coelho et al. 2021). We employed two complimentary metrics to quantify dynamic isoform changes: differential transcript expression (DTE) and differential transcript usage (DTU) (**Figure 6.4A**). For cases of DTE, we are measuring changes in overall transcript expression (counts per million, CPM) across conditions, identifying scenarios where the total expression of the isoform is changing (**Figure 6.4A**). For cases of DTU, we are measuring changes to the proportion of an isoform relative to others within the same gene, measuring how isoform usage shifts over time, even if the total expression remains constant (Soneson et al. 2016) (**Figure 6.4A**)

To ascribe a measure of statistical confidence in the isoforms that are differentially expressed across time, we used the samples with biological replicates (Day 0, Day 3, and Day 5). To identify cases of DTE occurring across these three timepoints, we used the time-series analysis function in the edgeR package (Robinson, McCarthy, and Smyth 2010). This analysis aided in identification of isoforms with dynamic changes across time.

To uncover specific patterns of isoform expression changes, we first filtered for isoforms exhibiting some statistically significant change in expression across any of the timepoints. We filtered for the top 500 isoforms identified at a false discovery threshold (FDR) less than 0.01 and performed hierarchical clustering on those expression values.

Two distinct clusters emerged. The members within cluster 1 represented isoforms whose expression gradually increased in concordance with EC development, while cluster 2 included isoforms with high expression that gradually decreased (**Figure 6.4B**). For genes in the identified clusters, we performed gene ontology (GO) enrichment, revealing that genes within these clusters generally correspond to developmental and metabolic pathways (*Appendix C: Supplemental Figure S3*).

We also performed time-series analysis on the gene-level expression patterns for Day 0, 3 and 5 to measure differential gene expression (DGE). We observed clustering patterns similar to those seen in the isoform expression results (*Appendix C: Supplemental Figure S4A*). Overall, the overlap of genes showing both DGE and DTE were low, as previously reported , suggesting that gene and splice regulation are orthogonal processes (*Appendix C: Supplemental Figure S4B*) (W. Li et al. 2016).

Next, we were interested in analyzing isoform changes for genes associated with EC identity. We used the PanglaoDB database (Franzén, Gan, and Björkegren 2019) to compile a list of genes that have been previously observed via single cell RNA-seq experiments in mouse and human, specifically expressed in EC populations (N=195). Overall, we found 38 EC-specific genes that had isoforms exhibiting DTE (*Appendix C: Supplemental Figure S5*). To specifically investigate isoform switching, we filtered the list to include genes with at least two isoforms, aiming to capture cases where one isoform's expression increases while another decreases, reflecting switches in ratios. Within this group, we found 28 EC-relevant genes expressing multiple isoforms and dramatically shifting in their fractional expression (**Figure 6.3C**).

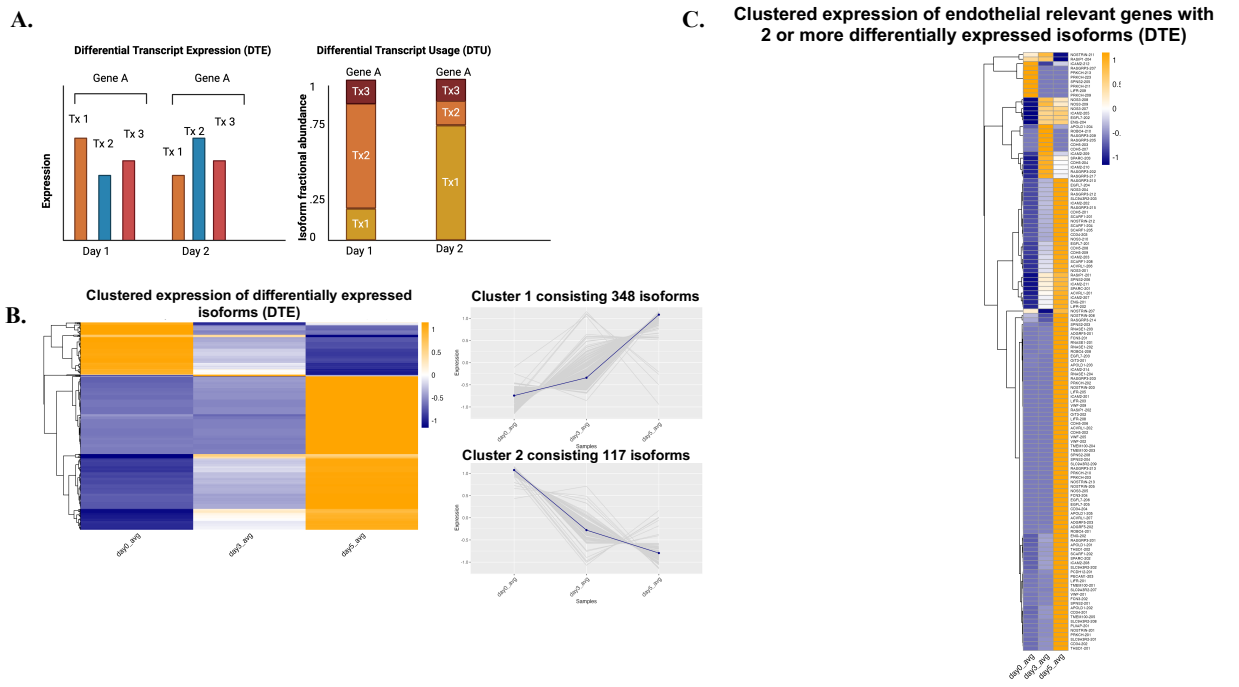


Figure 6.3 Dynamic isoform patterns observed during primordial EC differentiation.
A. Visualization of the difference between differential transcript expression (DTE) and differential transcript usage (DTU) cases used to profile isoforms systematically within this study B. Hierarchical clustering of the expression values (CPM) for the top 500 isoforms at FDR of less than 0.01 identified to have differential transcript expression profiles identified via dynamic expression analysis. C. Endothelial relevant genes with 2 or more isoforms identified to have DTE with their relative expression changes during differentiation

6.2.4 Differential transcript usage cases are more sensitive to genes with more isoforms

We next turned our attention to profiling isoforms that represented cases of DTU. Here we used spline fitting curves to model isoform expression changes across timepoints, to identify isoforms with statistically significant changes in their ratios (i.e., fractional abundance). We found 500 isoforms exhibiting DTU at a False Discovery Rate (FDR) threshold of <0.01. We found that this group of DTU isoforms profiled genes with several isoforms, suggesting this method is sensitive to minute changes in isoform expression fractions (*Appendix C: Supplemental Figure S6A*). Between the DTE and DTU analysis we found a small overlap (22 genes versus 416 genes in total for DTU and 381 in total for DTE) (*Appendix C: Supplemental Figure S6B*), suggesting these methods capture distinct regulatory mechanisms.

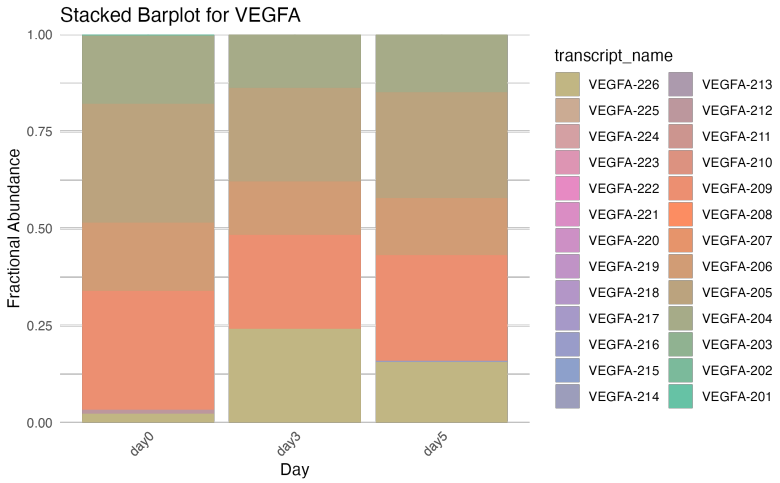
6.2.5 Visualizing the effects of isoform dynamics

For the isoforms we profiled, we wanted the ability to visualize how the usage of different isoforms relate to potential molecular effects, such as differential presentation of functional elements such as protein structural domains or active sites (Vitting-Seerup and Sandelin 2019). To achieve this, we utilized the transcripts collected via Kinnex as input to build

a custom University of Santa Cruz (UCSC) Genome Browser to enable visualization of the full-length transcripts identified and used existing genome features on the browser to associate regions on the collected transcript isoforms to corresponding protein domains. Obtaining dynamic vantage of transcripts within a one-dimensional view is difficult, however we focused on identifying the most highly expressed isoform for a gene (major isoform) on Day 0, 3, and 5 and color coded the major isoform for each day to observe isoform switching events (https://genome.ucsc.edu/s/emilyfwatts/madison). Such information can support hypotheses for how one isoform accomplishes a distinct function.

First, we looked at isoforms for the vascular endothelial growth factor A (*VEGF-A*) gene due to its well characterized splicing patterns that influence states of proangiogenic and anti-angiogenesis (Farrokh et al. 2015). Within our dataset we found proportions of isoforms changing (**Figure 6.4A**). When looking at the transcript structure we were able to see the dichotomy of isoforms denoted by the inclusion or lack of inclusion of exon 7 corresponding to the platelet derived growth factor domain responsible for accomplishing its distinct functions (White and Bix 2023) **(Figure 6.4B**).

A.



B.

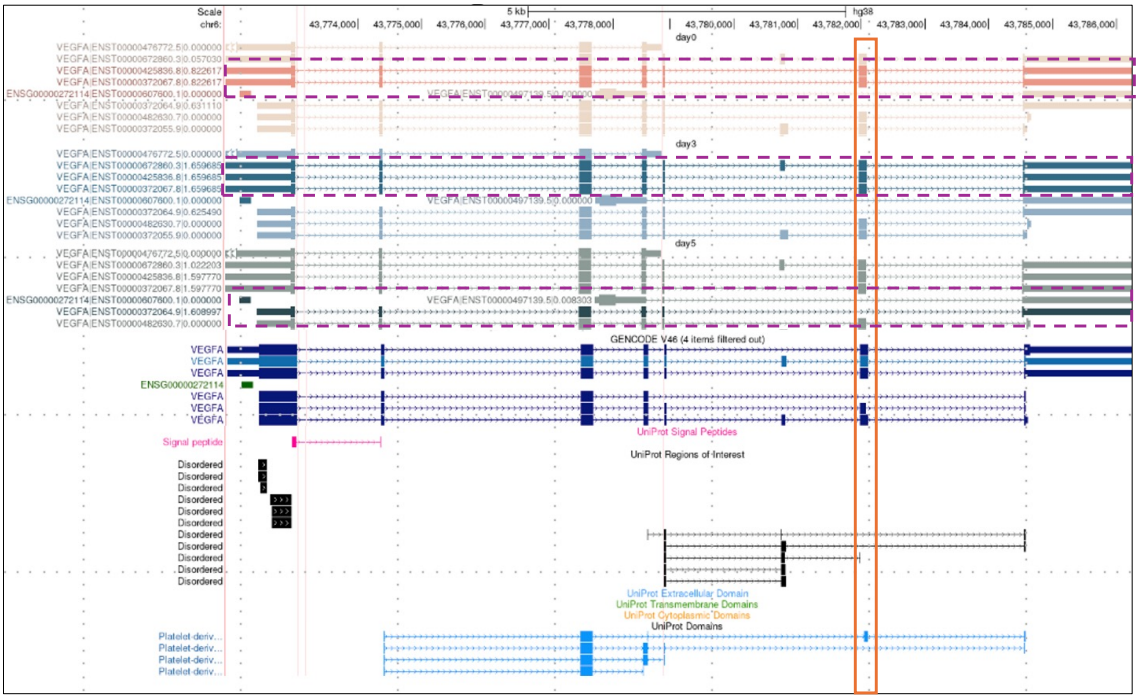
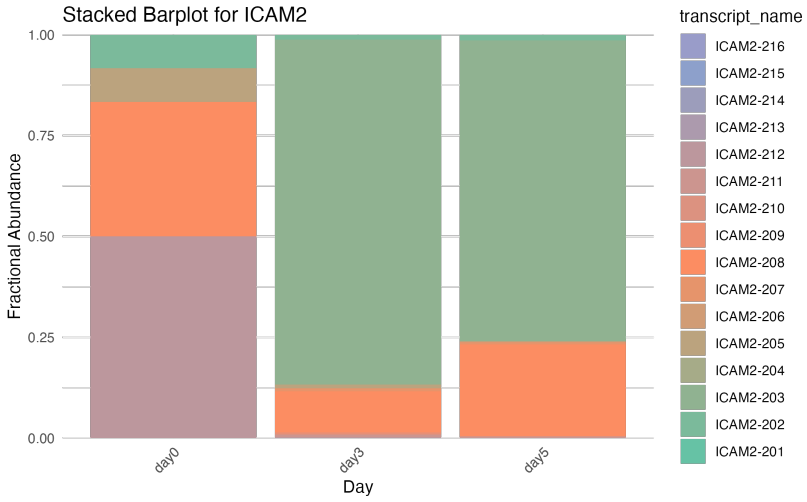


Figure 6.4 Dynamics of VEGF-A isoform expression
A. Stacked bar-plot of *VEGF-A* isoforms with respective ratios of isoforms over time B. UCSC genome browser track of the transcripts collected for *VEGF-A* from Kinnex sequencing, and the corresponding domains and protein features mapped to the transcript with the major isoforms highlighted in magenta and the region of the transcripts demonstrating the change in the platelet derived growth factor domain in orange (only a subset of transcripts shown for visualization purposes)

We next turned our attention to EC-relevant genes with isoforms having DTE, as these isoform switches represented robust changes in their expression. We looked at the gene *ICAM2,* noting its involvement with maintaining EC junctional integrity (Amsellem et al. 2014). We observed on Day 0 a shorter transcript inclusive of the signal peptide domain. On Days 3 and 5,

we observed the major isoform is modified to include an extracellular and immunogloloin domain (Guerra-Espinosa et al. 2024). We looked at the isoform expression patterns for *ICAM2* identified from our positive control from the purified EC Day 5 population observing similar trends in isoform ratios (*Appendix C: Supplemental Figure S7*).

A.



B.

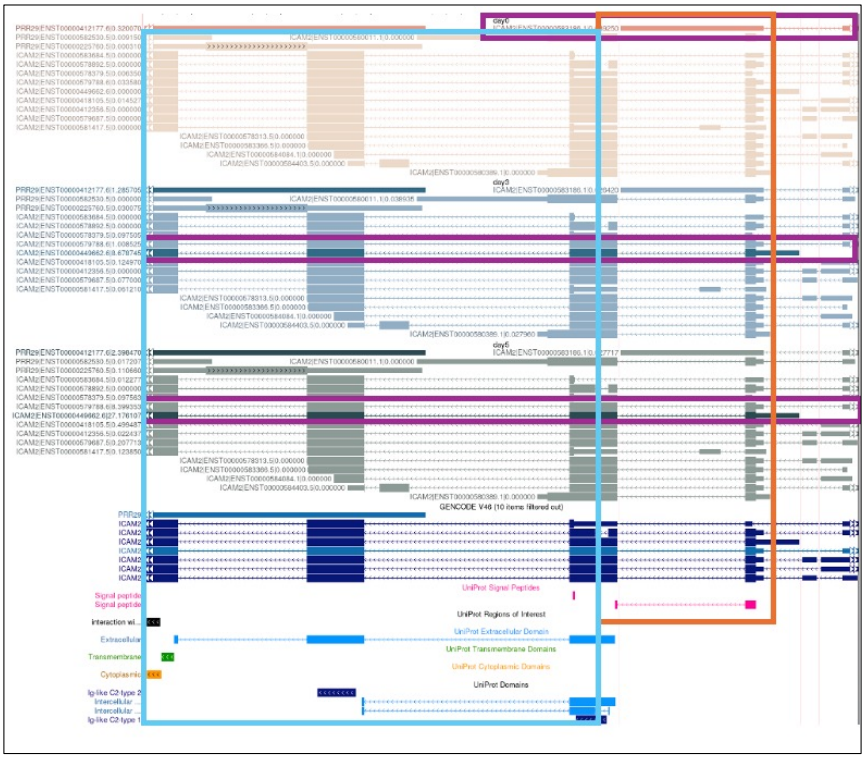


Figure 6.5 Differential transcript expression for the gene *ICAM2*

A. Stacked bar plot for *ICAM2* demonstrates dramatic switches in isoform ratios during differentiation. B. Browser track image of *ICAM2* transcripts identified demonstrating changes in the major isoform by day corresponding to changes within the signal peptide domain and the gain of multiple domains for the isoforms on Day 3 and Day 5.

6.2.6 Isoform dynamics of transcription factors and splice factors

Understanding that gene and isoform expression networks are greatly influenced by regulators working hierarchically, we focused on isoforms of transcription factors (TFs) and RNA-binding proteins (RBPs) found differentially expressed within the data. Overall we observed isoform dynamics associated with the serine and arginine rich splicing factor 5 (*SRSF5*), known to be involved in the splicing of the *VEGF-A* (Di Matteo et al. 2020) finding a large group of isoforms identified (Farrokh et al. 2015) (*Appendix C: Supplemental Figure S8*). We then turned to identifying dynamically regulated transcription factors. We observed the ETS transcription factor 2 (*ETV2)* to have DTE with isoforms differentially regulated between days (*Appendix C: Supplemental Figure S9*), noting the importance of this gene in regulating and promoting EC identity (Gong et al. 2022).

6.3 Discussion

Characterizing the landscape of isoforms involved in differentiation and development is a prerequisite towards understanding their contributions in modulating cell fate. To elucidate the isoform landscape associated with primordial EC differentiation, we sequenced and quantified isoform populations over the course of a six-day differentiation process from a pluripotent to EC population, collecting samples across time, identifying two groups of isoforms associated with EC differentiation. We identify a set of endothelial-relevant genes within this set, bringing attention towards the unexpectedly diverse isoform landscape in ECs, even for well-studied markers.

This dataset should serve as a resource for expressed isoforms underlying early EC differentiation and facilitate further functional studies to shed light on the importance of these isoforms in contributing to EC state. Functional testing could include *in vitro* studies to test for causal roles of an isoform, such as using molecular biology approaches to express or ablate certain isoforms and measuring its ability to promote or hinder differentiation. While this study has delineated the transcriptome, additional support for a transcript's biological relevance would be bolstered by protein-level expression evidence such as those provided via mass-spectrometry-based proteomics. Early methods we developed, such as "long-read proteogenomics" (**Chapter 2** and **Chapter 3**) should support these efforts.

6.4 Methods

6.4.1 Stem cell culture

Undifferentiated human induced pluripotent stem cells (hiPSCs): hiPSCs (WTC-11 iPSC line GM25256, NIGMS Repository number) were obtained from Coriell through the Materials Transfer Agreement with the University of Virginia from Dr. Hirschi's lab. WTC11 cells were thawed from CryoStor (Stem Cell Technologies, catalog #07932) per manufacturer's instructions and replated into Matrigel (Corning) basement membrane-coated wells with mTESR Plus (Stem Cell Technologies, catalog #100-0276) and ROCK (Stem Cell Technologies, catalog #73802).

WTC11 cells were cultivated utilizing mTESR Plus media on Matrigel coated plates with media changes following manufacturers' recommendations. Cells were cultured to maintain undifferentiated cell populations and passaged when confluent using ReLeSR (Stem Cell Technologies, catalog #100-0483) per manufacturer guidelines. Cells were replated at their desired density on Matrigel-coated dishes.

6.4.2 Stem cell (WTC11) derived primordial endothelial cells

WTC11 derived primordial endothelial cells: WTC11 cells were seeded with mTESR Plus media onto 6-well dishes or 10 cm Matrigel coated plates for primordial endothelial cell differentiation as described in Nelson et al., J Vis Exp 2021. Twenty-four hours after initial plating of WTC11 cells, on Day 0, media was aspirated and replaced with StemDiff APEL 2 (Stem Cell Technologies, catalog #05275) differentiation media containing 5μM GSK3i (CHIR99021, Reprocell, catalog #04-0004). On Day 1 media was aspirated and replaced with differentiation media containing 50ng/mL essential fibroblast growth factor (bFGF, R&D Systems, catalog #233FB025). For Days 3, 4, and 5 media was aspirated and replaced with differentiation media containing 25ng/mL of bone morphogenetic protein 4 (BMP4, R&D Systems, catalog #314BP010) and 50 ng/mL of vascular endothelial growth factor VEGF (Fisher Scientific). Cells were collected from Days 0-5 of the differentiation protocol using Accutase (ThermoFisher) per manufacturer guidelines and pelleted. Three biological replicates were obtained from Day 0 with Day 0-1 and Day 0-2 collected from a 10 cm dish and Day 0-3 from a 6-well dish. Three biological replicates were collected as described above for Day 5 with Day 5-1 and Day 5-2 derived from a 10 cm dish and Day 5-3 from a 6-well plate. Two biological replicates were obtained from Day 3 from 10 cm dishes. Days 1, 2 and 4 were all collected from 10 cm dishes. On average, for each sample, RNA was extracted from 1 X 10^6 cells using a RNeasy (Qiagen) kit with quality accessed via Bioanalyzer. In each RNA sample, Spike-In RNA Variants (SIRVs, Lexogen) were reconstituted per suggested manufacture guidelines and added to samples with Day 0 biological replicates receiving E1 SIRVS, Day 5 E2 and all other Days receiving S4. Master mixes of RNA and SIRVs were split into 4 tubes for parallel long-read PacBio Revio sequencing and short-read sequencing on the NovaSeq at 150bp.

6.4.3 qPCR gene marker analysis

Total RNA was isolated according to RNeasy (Qiagen, cat 74004) protocol from aliquots of 500,000 - 5 million cells. The concentration of the RNA was assessed by Nanodrop. RNA was converted to cDNA using the Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems protocol. RNA concentrations were normalized to the lowest concentration of RNA in batch.

After adding 2 μL of NEBNext RT primer mix, each sample was incubated at 72°C for 5 minutes. We combined NEBNext Single Cell RT Buffer, NEBNext Single Cell RT Enzyme Mix, and water in the following scheme to create a Master mix:

- NEBNext Single Cell RT Buffer - ((5 μL * 0.1) * # of RNA Samples)

- NEBNext Single Cell RT Enzyme Mix - ((2 µL * 0.1) * # of RNA Samples)
- Water - ((3 µL * 0.1) * # of RNA Samples)

The Master mix was vortexed, spun down, and was distributed in 10 µL aliquots to the samples after primary incubation. With 10 µL of the Master mix distributed to each sample and mixed, samples were incubated at 42°C for 1 hour and 15 minutes. After, 1 µL of NEBNext template switching oligo (TSO) was immediately added to each sample and mixed. Samples were incubated for another 15 minutes at 42°C.

For cDNA amplification, followed the Iso-Seq™ Express Template Preparation protocol based off the following scheme before distributing 80 µL of the resulting mix to each sample:

- NEBNext Single Cell cDNA PCR Master Mix - ((50 µL * 0.1) * # of RNA Samples)
- NEBNext Single Cell cDNA PCR Primer - ((2 µL * 0.1) * # of RNA Samples)
- Water - (28 µL * (# of RNA Samples * 0.1)

Primary mixture was distributed to each sample and vortexed then placed in a thermal cycler to begin the Iso-Seq™ Express Template Preparation protocol. The PCR included an initial denature step at 98°C for 45 sec, 14 cycles of denaturation at 98°C for 10 sec, annealing at 62°C for 15 sec, and extension at 72°C for 3min. Final Extension was at 72°C for 5min.

cDNA was cleaned via ProNex Size-Selective Chemistry (Promega, cat NG2001). Samples were eluted with 32.5µL of Elution Buffer (Promega, cat NG116A). Concentration of cDNA was measured on Qubit (Invitrogen, dsDNA High Sensitivity).

qPCR was run on a QuantStudio 6 Real-Time PCR system (Applied BioSystems catalog #4485691). First, we prepared a stock of primers: 5 µL of 1x Power-up SYBR Green Mastermix (ThermoFisher, catalog #A25741) and the second stock of cDNA was 5 µL of 1x Power-up SYBR Green Mastermix that contained 38 ng of cDNA. The 2 stocks were mixed together in MicroAmp Optical 384-well plates (Applied Biosystems, catalog # 4309849), centrifuged, and sealed.

The final mixture contained 3.33ng of cDNA per reaction, 1x Power-up SYBR Green Mastermix (ThermoFisher, catalog #A25741), 0.5uM of each primer in 10 µL of total volume. Amplification began on the thermal cycler. Collected data was analyzed and normalized according to the GAPDH sample run in the experiments as a control. Then we normalized every sample from different time points to Day 0 using delta CT (ΔΔCT). Fold change was calculated using 2^(–ΔΔCT).

6.4.4 SIRV preparation and sample aliquoting

Extracted total RNA was RNA was analyzed on an Agilent Bioanalyzer to confirm RNA integrity for downstream analysis. We consistently observed an RNA integrity value (RIN) value of ~10 for all samples. After assessment of quality via Bioanalyzer, Spike-In RNA Variants (SIRVs, Lexogen, catalog #025.03 and 025.03) were prepared and reconstituted per suggested

manufacture guidelines and added to samples with Day 0 biological replicates receiving E1 SIRVS, Day5 E2 and all other Days receiving S4. Master mixes of RNA and SIRVs were split into 4 tubes for parallel long-read PacBio Revio sequencing and short-read sequencing on the NovaSeq at 150bp.

6.4.5 Short-read RNA sequencing

Aliquots were obtained as described above from the master mixes made with the total RNA and SIRV spike-ins. Kapa RNA HyperPro Kits (catalog # 08098093702) were used for RNA sequencing library preparation. Resulting libraries were sequenced on the NovaSeq at 150bp.

6.4.6 Kinnex long-read RNA sequencing library preparation and long-read RNA sequencing run

PacBio (Kinnex) data were collected from an aliquot of the extracted total RNA collected from the SIRV containing master mixes as described above. From this RNA, cDNA was synthesized using the Iso-Seq Express kit (PacBio). Approximately 300 ng of cDNA from each sample was barcoded and equally pooled. Samples were subjected to Kinnex multiplexing to generate Kinnex concatemers for final SMRTbell molecule generations. Resulting libraries were sequenced on the Revio.

The samples were demultiplexed bioinformatically using pb-demux and processed using Pigeon. The intermediate files (mapped.bam) from Iso-Seq were used for Bambu processing.

6.4.7 Purified Day 5 primordial endothelial cell long-read RNA-sequencing data collection

WTC11 cells were subjected to the primordial endothelial cell differentiation described above. Cells were harvested on Day 5 and were sorted using the FACS Melody sorter to select for cell populations that were CD31+CD45- FACS sorted endothelial cells. These cells were then processed via the PacBio Iso-Seq protocol and subjected to sequencing on the Sequel IIe capturing ~ 3 million CCS reads. Resulting files were analyzed via Iso-Seq to generate the distinct set of transcripts and genes.

6.4.8 Transcript isoform classification

BAM files were generated via Iso-Seq from the Kinnex-derived long-read sequencing data and were used as input into Bambu to generate the set of distinct full-length transcripts and associated genes. Additional classification was performed via SQANTI3 version 5.2 to annotate Bambu generated GTF files. The inputs for SQANTI3 analysis include the GENCODE version 35 annotations (i.e., GTF file) and the human reference genome (GRCh38, only canonical chromosomes chr1-22, X, Y) as well as the Bambu generated GTF file. The SQANTI3 outputs, isoform and junction “classification” files were subjected to additional analysis using custom R scripts.

6.4.9 Generation of UCSC genome browser tracks

We used SQANTI to generate a CDS (file of the genomic regions) of the genes and isoforms quantified via Bambu. Common gene names were replaced using a custom script (https://github.com/efwatts/LRP_Troubleshooting/tree/main/15_accession_mapping). In order to assess changes in isoform expression we generated 3 separate GTF files to denote the most abundant isoform ("major isoform") for a gene at Day 0, 3, and 5 and added associated color shading based off CPM. Both the GTF files for each day and associated CDS files were uploaded to the UCSC genome browser using custom tracks.

6.4.10 Data analysis and plot generation

All downstream analysis was performed via custom R scripts and Bioconductor available packages. Custom scripts were made via R and can be found at this repository https://github.com/sheynkman-lab/madison_mmehlferber/tree/main/scripts.

6.4.11 Availability of materials

The RNA-sequencing data can be found at https://downloads.pacbcloud.com/public/dataset/MAS-Seq-bulk-2023-WTC11/. All other data files used during this analysis can be found at Box (https://virginia.box.com/s/9zs2kntwppncozjuvyzofixky419png5).

Chapter 7 Concluding remarks and future directions

The field of isoform biology is an exciting and rapidly evolving research area. From the discovery of isoforms spearheaded by Phil Sharp and colleagues in 1970, to tracking splice event expression with RNA-sequencing technologies, and most recently, associating specific isoform aberrations as indicators of disease, the field has seen blossoming advances (Berget, Moore, and Sharp 1977; Z. Wang, Gerstein, and Snyder 2009; Cooper, Wan, and Dreyfuss 2009). A landmark finding in this field was the discovery in 2008, as evidenced through RNA-sequencing approaches, that over 95% of human protein coding genes undergo the process of alternative splicing to yield multiple isoforms with potentially distinct functions (E. T. Wang et al. 2008; Pan et al. 2008). Since this discovery, researchers have continued to inform isoform discovery with recent predictions indicating that there are at least 100,000 alternative transcripts, with active research increasing this number (Jiang and Chen 2021). Here, I describe my contributions to the isoform biology field which included improving the ability to profile and resolve isoforms and their associated protein isoforms. I applied a long-read proteogenomic method to construct an isoform atlas within ECs, recognizing that such knowledge could elucidate the intricacies of EC identity, enabling the diverse and critical functions within the cardiovascular system.

I first discussed my contributions in developing an integrated approach to improve identification of isoforms and linking corresponding protein-level evidence for such isoforms through the development of the long-read proteogenomics method in **Chapter 2**. We leveraged the resolution of long-read RNA-sequencing to detect intricate AS events, overcoming limitations of short-read RNA-sequencing for isoform discovery. In tandem with mass-spectrometry proteomics, we created an open-source computational pipeline to generate sample-specific isoform databases to track isoforms at both the transcriptome and proteome level. As part of this work, I demonstrated the discovery of novel protein isoforms by identifying peptides mapping to AS regions of novel transcripts, providing protein-level evidence of novel protein isoform events.

In **Chapter 3**, I applied this integrated long-read proteogenomic (LRP) method onto human umbilical vein endothelial cells (HUVECs) demonstrating the versatility of the LRP approach towards characterizing the isoform landscape in different tissue types. This work represented the first application of the approach in supporting isoform discovery in ECs. Despite the well-studied status of HUVECs due to their experimental tractability to model vascular phenotypes, I documented novel protein isoform events for key genetic markers of EC identity, including *CDH5* and *PECAM1*. This work demonstrated the ability of the long-read proteogenomic approach to profile unique splicing patterns, even for very well characterized genes.

In **Chapter 4**, I reviewed how isoforms derived from AS change during cell development. I explored the splice repertoire of isoforms that govern distinct developmental states as well as their involvement in modulating both gene and splice regulatory networks, describing the resulting transcriptional and splicing networks they mediate. I concluded that cell fate regulation is deeply interconnected within an integrated network and that further studies delineating the factors involved in development is important to understand how a cell arrives at a final cell state.

With the knowledge that the use of a particular isoform can dictate cell fate, I wanted to characterize isoforms involved in the establishment of primordial ECs, the precursors to mature ECs. I utilized a tractable induced pluripotent stem cell system to model their development and collected time course RNA-sequencing data so I could interrogate day-to-day changes in global isoform expression. Capitalizing on increases in throughput enabled by PacBio long-read RNA-sequencing, I assessed this platform for the purpose of multi-sample isoform characterization in **Chapter 5**. I described bioinformatic approaches for analyzing multiple sequencing datasets, extending isoforms studies from single samples to multi-sample dynamics.

Encouraged by the ability of this new technology to further enhance isoform characterization, I utilized this technology in **Chapter 6** to describe isoform dynamics during the differentiation of primordial ECs. I profiled various approaches to capture changes in isoform expression patterns temporally. I highlight endothelial-relevant genes having dynamically expressed isoforms. I provided an approach to visualize the structure of isoforms and those changing over the time course. While this study illuminated the extensive isoform landscape associated with primordial EC development, associating a specific function to an isoform remains an ongoing effort. However, the study serves as an atlas representing the rich isoform landscape present in early ECs.

7.1 Future Directions:

As RNA-sequencing technologies continue to evolve, resolution of the complex isoform landscape within the human genome will continue to advance. Recently long-read RNA-sequencing was chosen as the method of the year (“Method of the Year 2022: Long-Read Sequencing” 2023). With the knowledge that isoforms operate in tissue specific manners, open source databases such as GTEx provide a resource of various tissue transcriptomes (GTEx Consortium 2013). However, databases highlighting temporal transcriptome changes associated with arrival at a cell fate remain limited. Recognizing that the data discussed in this thesis primarily surrounds bulk transcriptomic datasets, it is important to note that the evolution of single-cell long-read RNA-sequencing offered through PacBio, offers isoform resolution at individual tissue level.

As the cost of sequencing throughput continues to decrease, knowledge of the complexities of the human genome will continue to be elucidated. With the sheer number of isoforms reported in long-read RNA-sequencing studies, the question shifts to the plausibility of their existence and associated function (Weirick et al. 2016). Recent studies have suggested that AS provides

regulation by creating unproductive transcripts or transcripts that undergo nonsense mediated decay (NMD). Such events suggest a limited role of AS in contributing diversity to the proteome (Fair et al. 2024). Therefore, efforts to comprehensively profile the proteome to parse the relationship of the transcriptome to the proteome will remain important. Mass-spectrometry based proteomics remains the gold standard approach. However, traditional MS data dependent workflows return limited numbers of isoform specific peptides to support the existence of protein isoforms (Korchak et al. 2024). To increase versatility of mass-spectrometry for such applications, targeted mass-spectrometry strategies have gained traction. In these advanced targeting methods, candidate peptides of interests are identified, representing potentially novel regions, and targeted by the mass spectrometer to acquire spectra from the peptides of interest (Reed et al. 2022; Yu et al. 2020; Stopfer et al. 2021). This approach can be used to isolate and confirm the presence of isoform-specific peptides. Another approach that might support protein detection is single molecule protein-sequencing (Alfaro et al. 2021). This method does not provide global protein sequence coverage but rather focuses on individual proteins, identifying their amino acid sequences (Floyd and Marcotte 2022).

However, researchers continue to confound the activity of detecting an isoform and associating a function. These issues are not trivial, and very few studies currently exist offering this full circle perspective globally (Shaw et al. 2022). Within the isoform field there is this natural desire to utilize the power of isoforms to support precision medicine techniques inspired by the knowledge that aberrant splice patterns can separate healthy and diseased tissues. Comprehensive studies illustrating the direct role of individual isoforms are needed for the isoform biology field to demonstrate its applicability to precision medicine.

Early efforts are beginning to screen isoform functions at high-throughput. Studies such as those employed by Dr. Julia Joung et al. (Joung et al. 2023) cloned libraries of transcription factors into stem cells to create an expression atlas associating specific roles of transcription factors with cell phenotypes. Advancements in gene-editing platforms such as Cas13 provide precise modulation of RNA molecules, and this could be used to edit isoforms (Abudayyeh et al. 2017).

It is important to note the expense and resource-intensive nature of isoform-resolved panel studies. Supplementing these experimental efforts, computational approaches can predict functions of proteins with high accuracy (Mishra, Muthye, and Kandoi 2020). Tools such as AlphaFold, winning the 2024 Nobel Prize (Callaway 2024) have been applied to predict accurate protein structures informing how a transcript structure could alter a protein domain (Abramson et al. 2024). Thus, predictive tools could alleviate the burden of resource-intensive experimental approaches to study isoforms. Additional computational tools, such as isoform functional annotation databases, are gaining traction allowing researchers to associate a genetic pathway with an isoform (Ferrer-Bonsoms et al. 2020).

These studies emphasize the need to embrace systems-level perspectives to connect the multitude of networks contributing to a biological system to draw true meaning (Hood et al. 2004). With advances in technology comes the responsibility to utilize it effectively to support and inform the biomedical community. Therefore, for the isoform biology field to reach its full potential requires collaboration between the technology development, basic science and the clinical community. Welcoming these approaches could yield breakthroughs in the field of precision medicine. With this attitude, I am optimistic that an isoform might be discovered and utilized as the therapeutic driver to treat a patient cohort in the next decades.

Finally good science comes not only from the impact one has on the biomedical community but the ways in which one leads that team to discovery of approaches that take the leap from the bench to the bedside. Strides emphasizing reproducible and open access science is the key to contributing to long-lasting and impactful medical breakthroughs (M. D. Wilkinson et al. 2016).

In conclusion, I detail here my efforts related to better profiling isoform landscapes, especially as it relates to contributing to the EC atlas. I illustrate the benefits of a sample matched database for isoform discovery highlighting the ability to profile novel protein isoforms for previously uncharacterized splicing events. I demonstrate the applicability of utilizing time series RNA-sequencing experiments to study dynamic isoform populations and offer an approach to effectively analyze multidimensional sequencing data. And lastly, I hope one can appreciate the development of a comprehensive method of profiling isoforms with great versatility and utility for the larger biomedical community.

Appendix A: Supplemental Figures for Chapter 3

The supplemental figures here pertain to the data within **Chapter 3**.

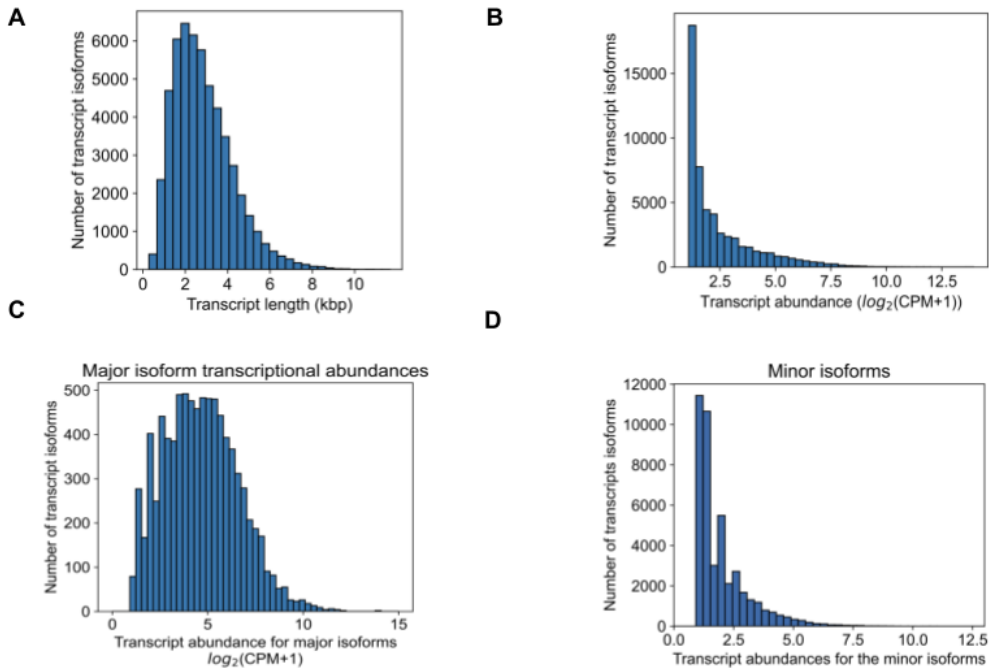


Figure S1. Characterization of the HUVEC full-length transcriptome based on long-read RNA-seq data. (A**)** Distribution of transcript isoform length (B) Distribution of the transcript abundance (counts per million, CPM) (B) Distribution of all transcripts identified from the Iso-Seq pipeline with abundance greater than one CPM. **(**C) Distribution of the abundances for the most highly expressed isoform for each gene, i.e., the major isoform. (D) Distribution of the abundances of the minor isoforms reported for all transcripts with CPM>1.

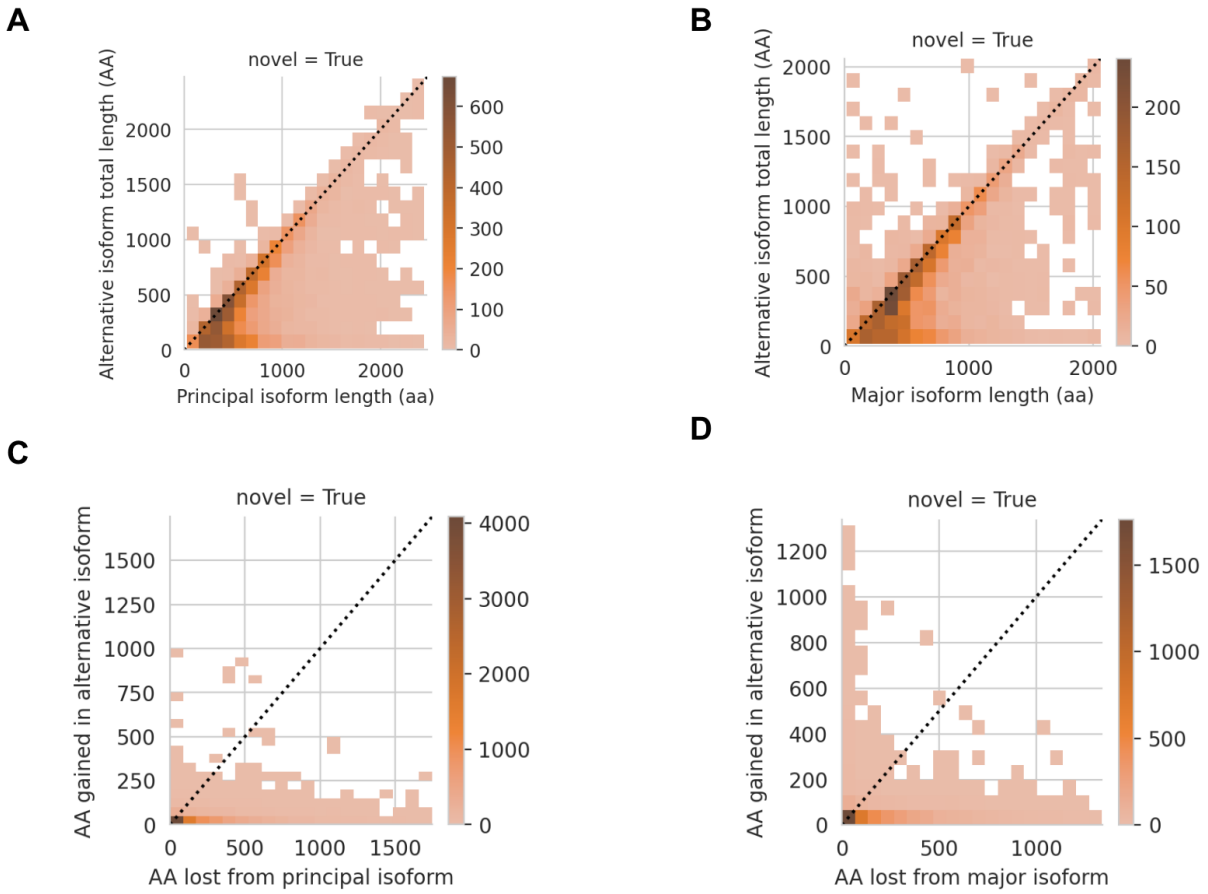


Figure S2. Characterization of novel isoform length as identified through long-read proteogenomics (A) Overall length in amino acids of the principal isoforms as defined by APPRIS (B) Overall length in amino acids of the major isoforms as defined within HUVECs (C) Comparison of the length in amino acids of identified novel isoforms (novel in catalog (NIC) and novel not in catalog (NNC)) against the APPRIS reference isoform. (D) Comparison of the length in amino acids of identified novel isoforms (novel in catalog (NIC) and novel not in catalog (NNC) against the "major isoform" reference isoform as identified within HUVECs.

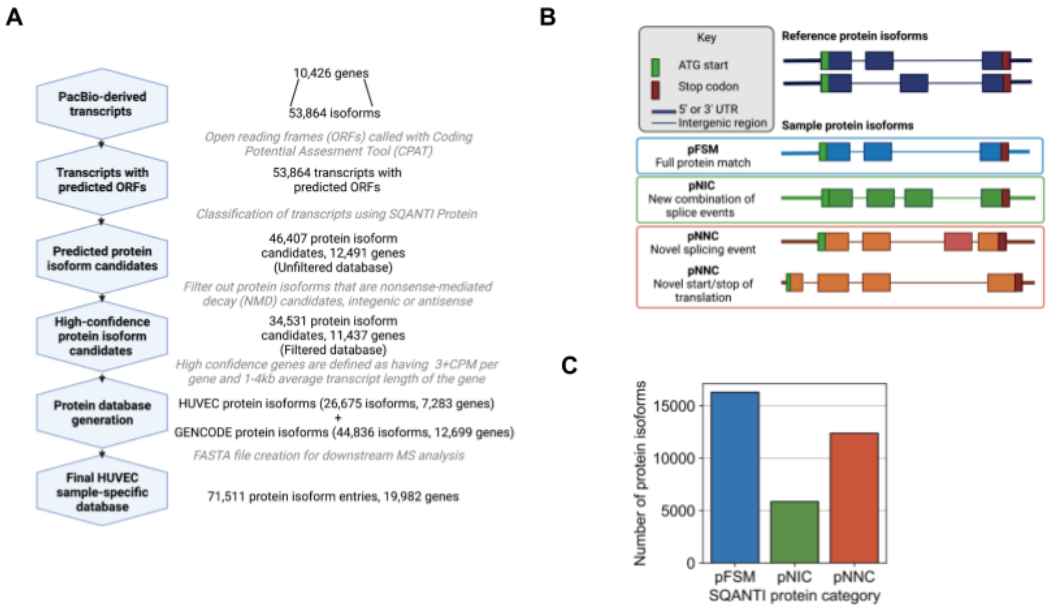


Figure S3. Derivation of predicted protein isoforms to generate a HUVEC sample-specific database. (A) Schematic of protein database generation from long-read RNA-seq data. (B) Schematic of SQANTI Protein classification (C) Bar chart indicating the frequency of different protein isoforms based on novelty category.

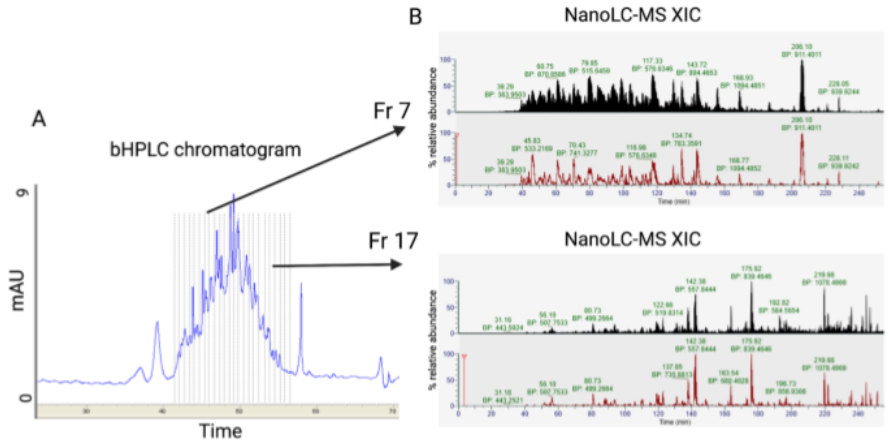


Figure S4. Basic pH HPLC fractions for the HUVEC MS data collection. (A) UV trace of peptide elution from offline fractionation (range of fractions shown with dashed lines). (B) Total ion (black) and base peak (red) chromatograms from LC-MS analysis of peptides from representative fractions 7 and 17. Basic High Performance Liquid Chromatography (bHPLC), Extracted ion chromatogram (XIC), Nanoscale liquid chromatography coupled to tandem mass spectrometry (nano LC–MS/MS), milli-Absorbance Units (mAU).

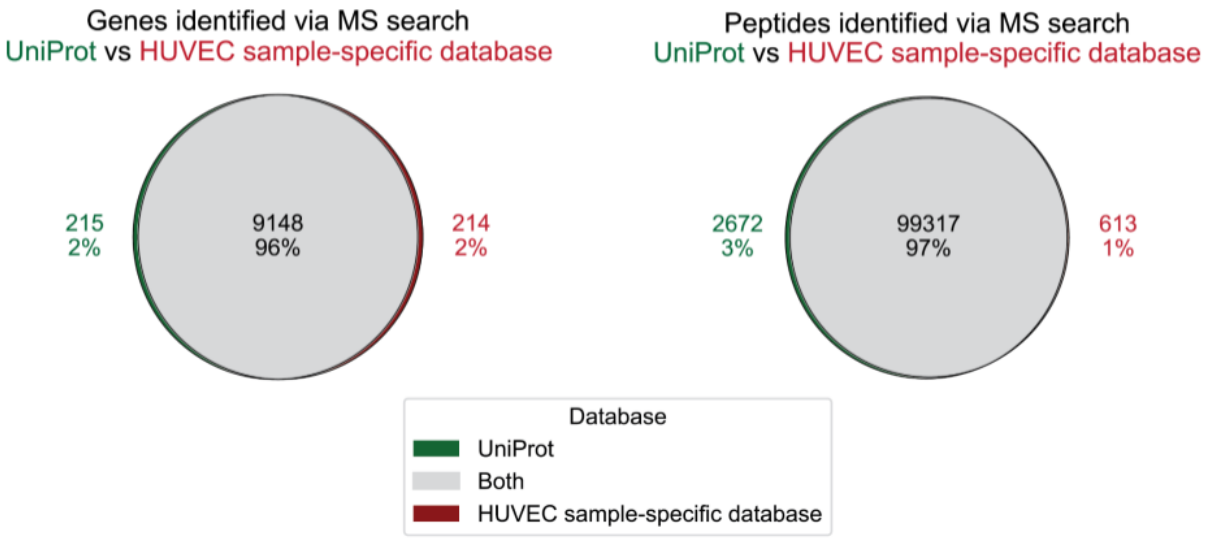


Figure S5. Comparison of proteomic coverage when using the HUVEC sample-specific versus UniProt protein databases for MS searching.

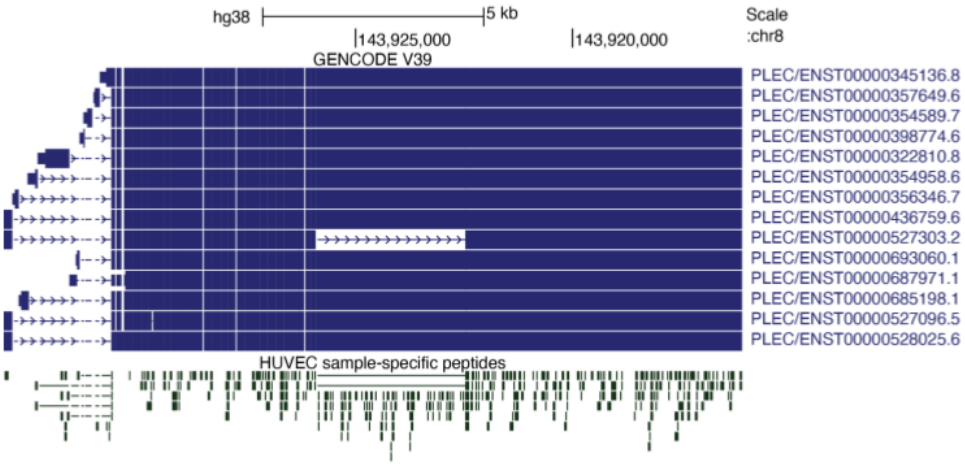


Figure S6. Plectin (*PLEC*) gene evidenced by seven unique isoforms. Plectin (*PLEC*) gene in which MS analysis identified seven unique isoforms each evidenced by their own unique peptide.

Appendix B: Supplemental Tables for Chapter 3

The supplemental tables here pertain to the data in **Chapter 3** and can be found at: https://www.tandfonline.com/doi/suppl/10.1080/15476286.2022.2141938?scroll=top

Appendix C: Supplemental figures for Chapter 6

The supplemental figures here pertain to the data within **Chapter 6**.

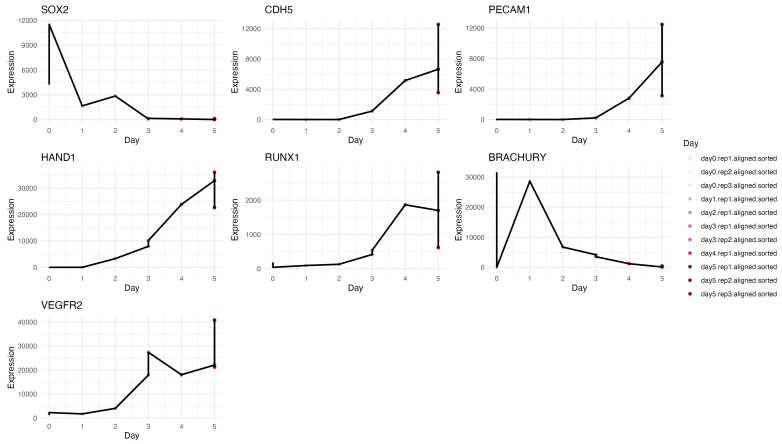


Figure S1: In silico-derived qPCR of genetic markers associated with cell fate markers from long-read RNA-sequencing derived gene matrix
Long read RNA-seq quantification results for genetic markers associated with cell fate transitions. There is a decrease in pluripotency genes (*SOX2*) accompanied by an increase in expression of early mesoderm genes (*BRACHYURY*) and mesoderm (*HAND1*) followed by upregulation of genes associated with early endothelia (*CDH5*, *VEGFR2*).

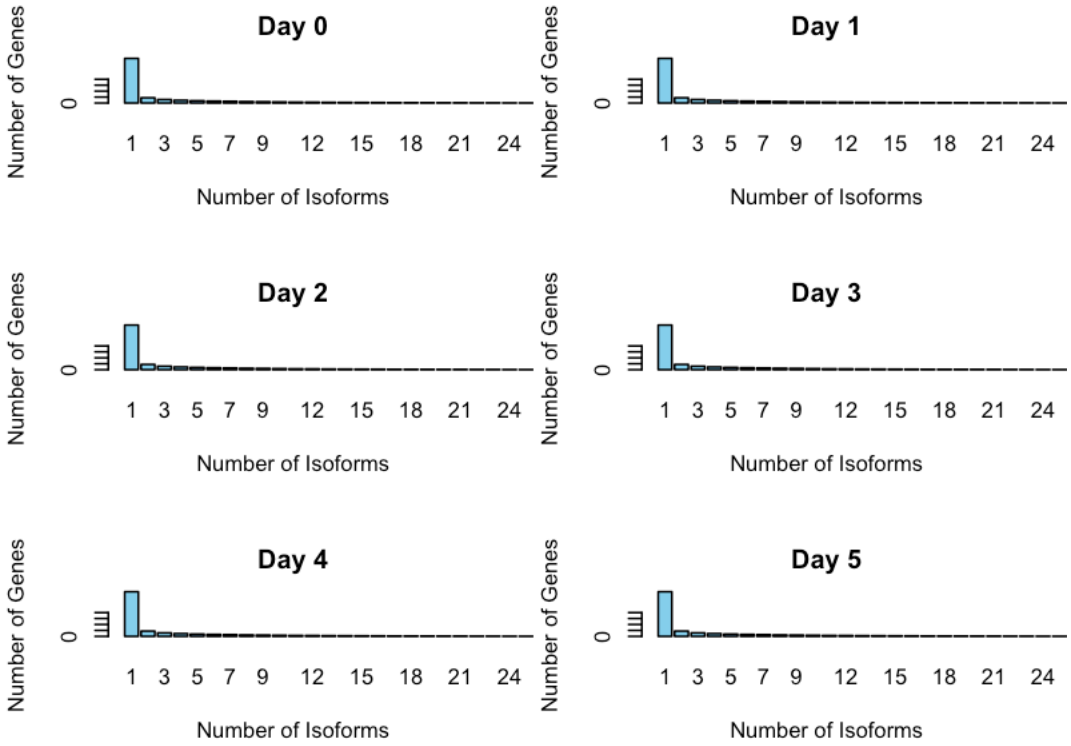


Figure S2: Number of genes expressing multiple isoforms by Day

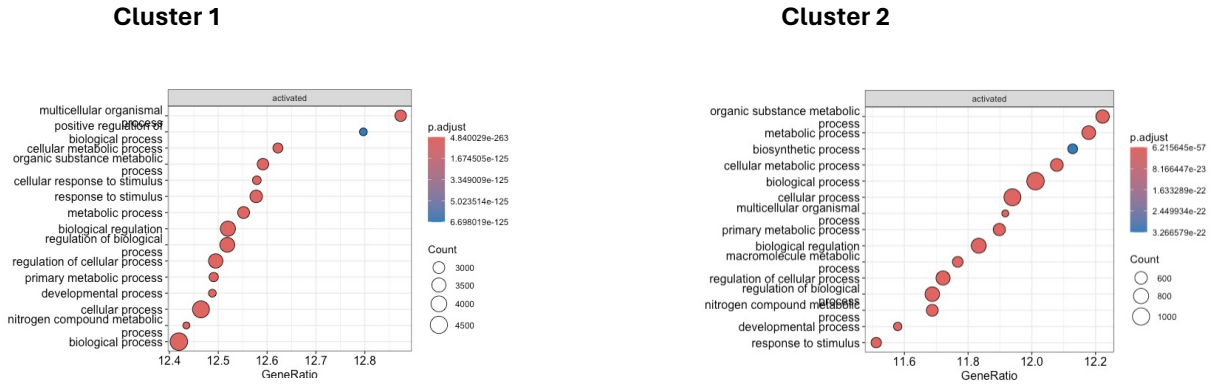


Figure S3: GO enrichments for the parental genes with their isoforms having DTE. Gene ontology of the genes comprising the two clusters of genes with isoforms having differential transcript expression.

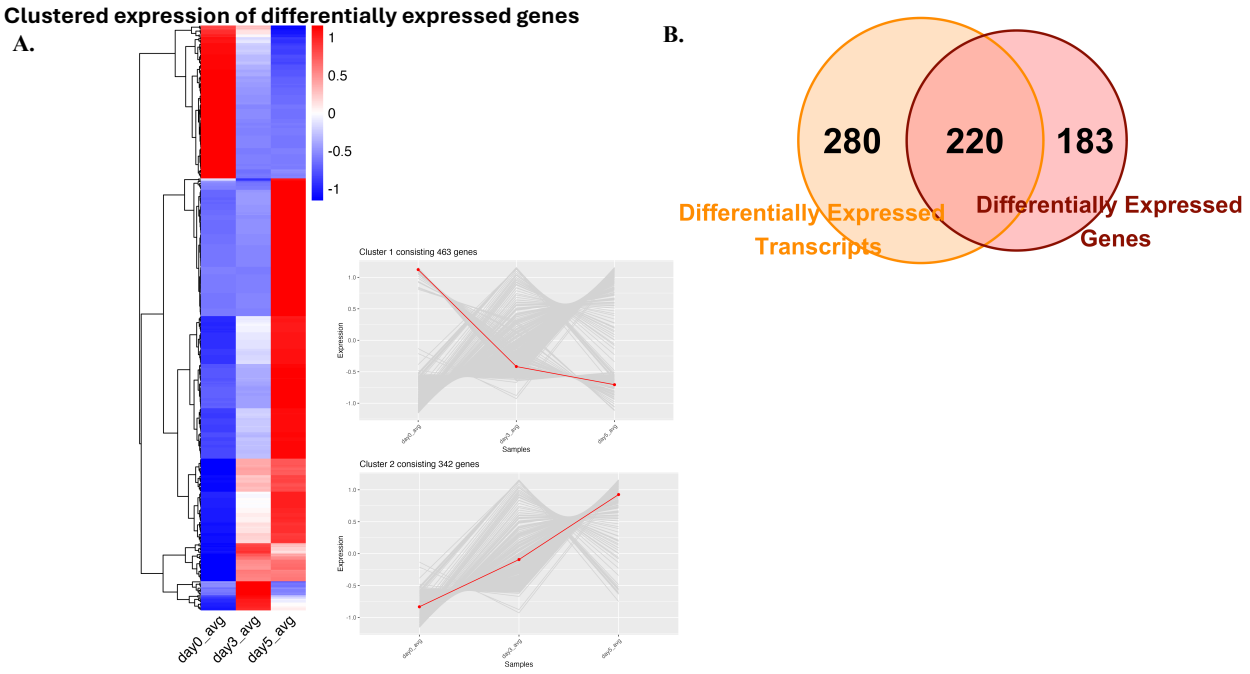


Figure S4: Gene dynamics and clustering. A. Clustering of gene-level expression profiles for those genes with differential expression over the time course B. Overlap of the genes having differential gene expression (DGE) and the genes with isoforms having differential transcript expression (DTE).

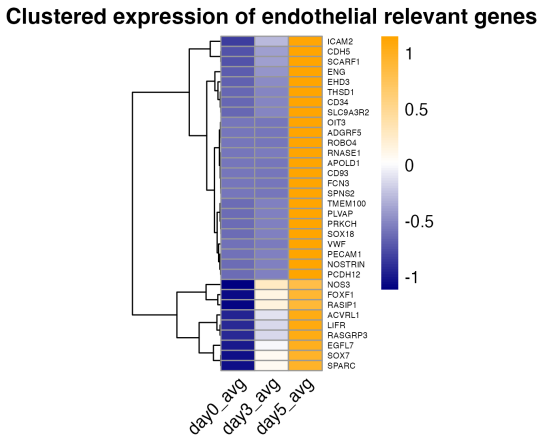


Figure S5: Endothelial-relevant genes with isoforms classified with differential transcript expression (DTE). Gene-level expression changes for the DTE EC-relevant genes.

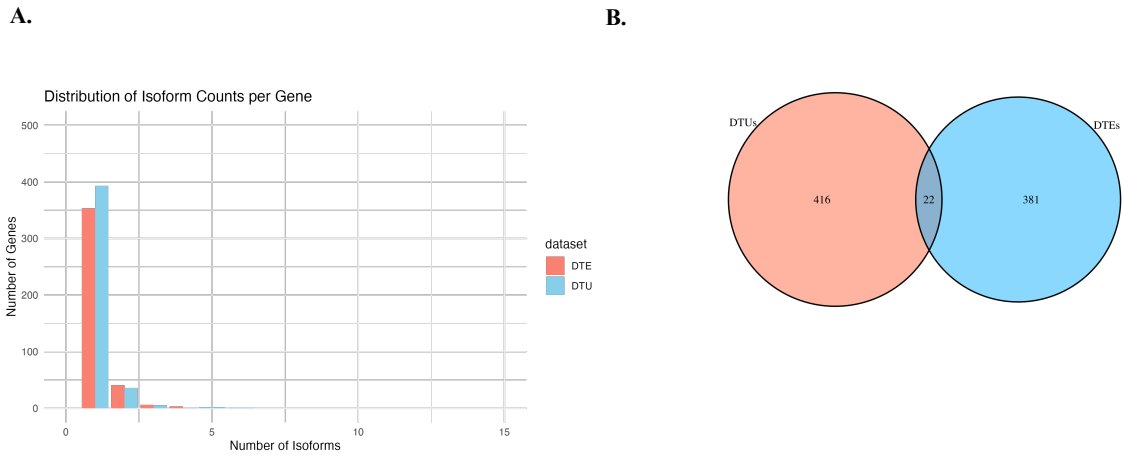


Figure S6: Number of isoforms profiled between differential transcript expression (DTE) and differential transcript usage (DTU) approaches. A. Number of genes with multiple isoforms for DTE vs. DTU approaches B. Overlap of genes exhibiting DTE vs. DTU.

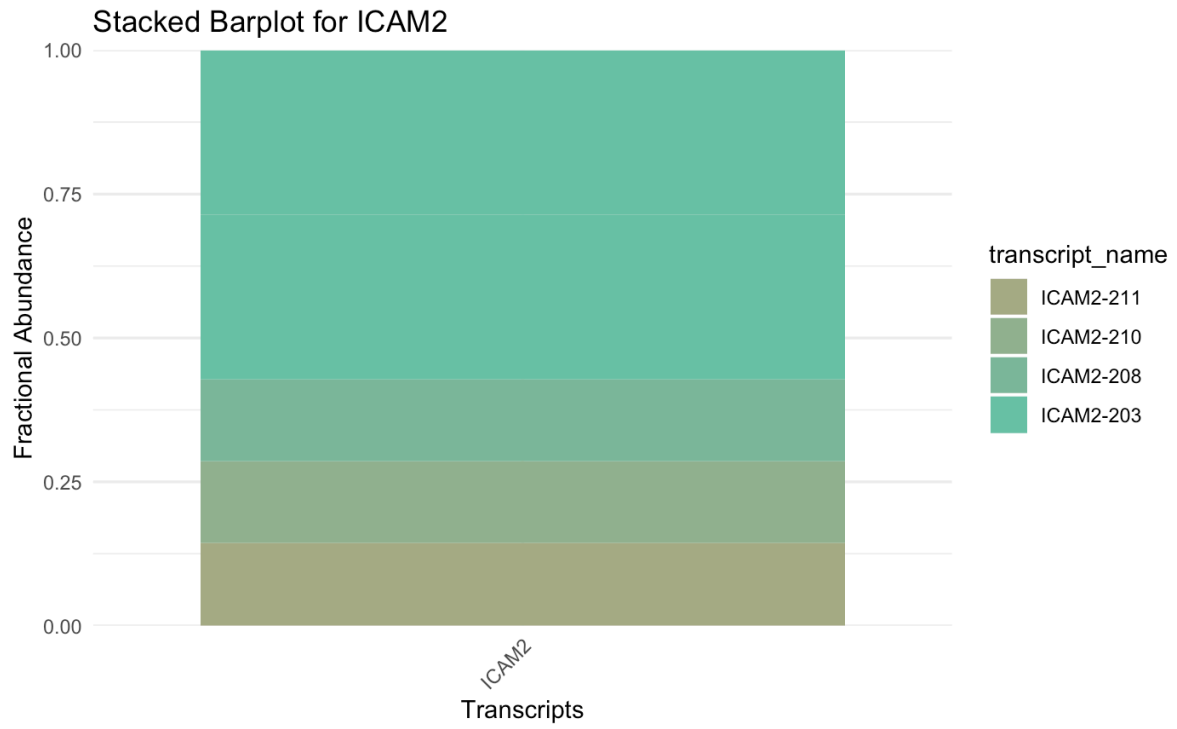


Figure S7: Purified primordial endothelial cell Day 5 isoform ratios for the gene *ICAM2*.

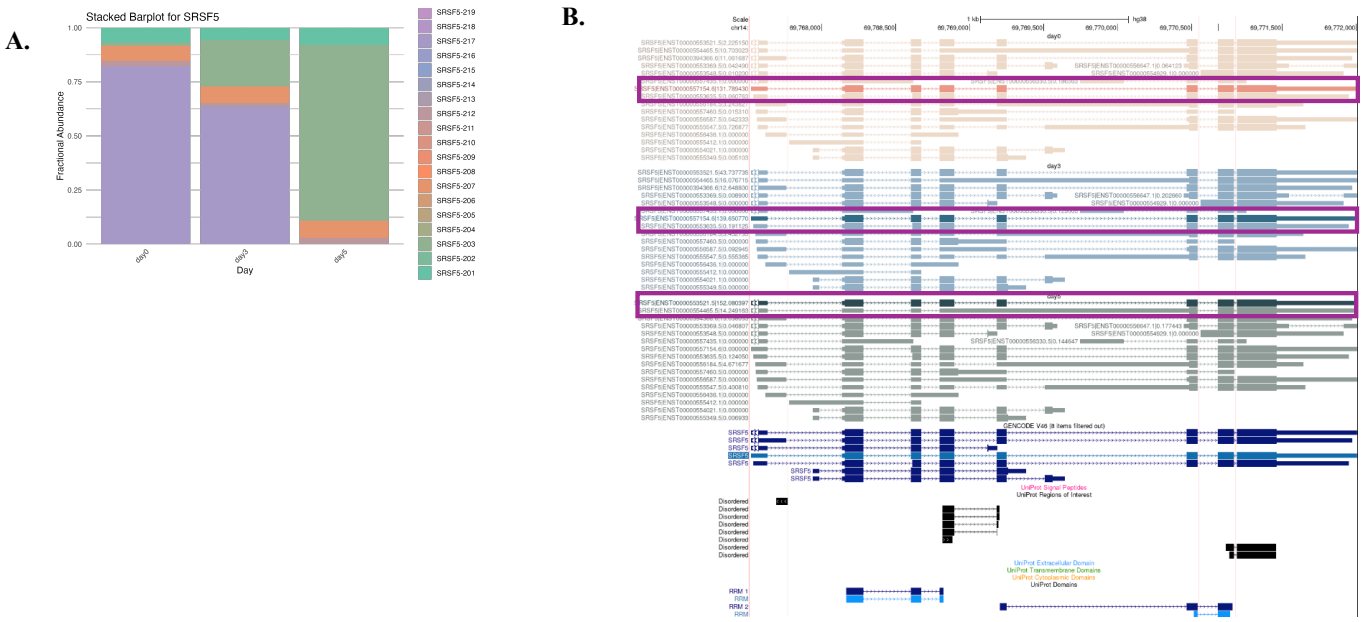


Figure S8: Isoform dynamics for the splice factor *SRSF5* in ECs. A. The splicing factor, *SRSF5*, regulates *VEGF-A* isoforms. It is found to have dynamic isoform expression changes observed over differentiation. B. Browser track for *SRSF5* isoforms with the major isoforms outlined in magenta.

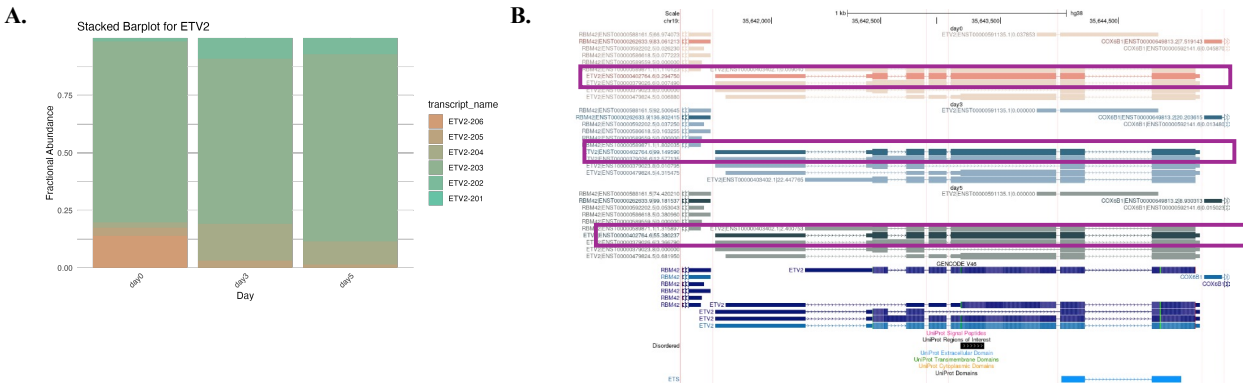


Figure S9: Isoform dynamics in transcription factor *ETV2* in ECs. A. Stacked bar plot for the gene *ETV2*, which is a known regulator that promotes endothelial cell function B. UCSC genome browser highlighting the major isoform for *ETV2* outlined in magenta for each day. Noting the reduction in expression of ETV2-206 during differentiation.

References:

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, et al. 2024. “Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3.” *Nature* 630 (8016): 493–500.

Abudayyeh, Omar O., Jonathan S. Gootenberg, Patrick Essletzbichler, Shuo Han, Julia Joung, Joseph J. Belanto, Vanessa Verdine, et al. 2017. “RNA Targeting with CRISPR-Cas13.” *Nature* 550 (7675): 280–84.

Aebersold, Ruedi, Jeffrey N. Agar, I. Jonathan Amster, Mark S. Baker, Carolyn R. Bertozzi, Emily S. Boja, Catherine E. Costello, et al. 2018. “How Many Human Proteoforms Are There?” *Nature Chemical Biology* 14: 206.

Agirre, E., A. J. Oldfield, N. Bellora, A. Segelle, and R. F. Luco. 2021. “Splicing-Associated Chromatin Signatures: A Combinatorial and Position-Dependent Role for Histone Marks in Splicing Definition.” *Nature Communications* 12 (1): 682.

Aguzzoli Heberle, Bernardo, J. Anthony Brandon, Madeline L. Page, Kayla A. Nations, Ketsile I. Dikobe, Brendan J. White, Lacey A. Gordon, et al. 2024. “Mapping Medically Relevant RNA Isoform Diversity in the Aged Human Frontal Cortex with Deep Long-Read RNA-Seq.” *Nature Biotechnology*, May. https://doi.org/10.1038/s41587-024-02245-9.

Alamancos, Gael P., Amadís Pagès, Juan L. Trincado, Nicolás Bellora, and Eduardo Eyras. 2015. “Leveraging Transcript Quantification for Fast Computation of Alternative Splicing Profiles.” *RNA* 21 (9): 1521–31.

Alfaro, Javier Antonio, Peggy Bohländer, Mingjie Dai, Mike Filius, Cecil J. Howard, Xander F. van Kooten, Shilo Ohayon, et al. 2021. “The Emerging Landscape of Single-Molecule Protein Sequencing Technologies.” *Nature Methods* 18 (6): 604–17.

Al’Khafaji, Aziz M., Jonathan T. Smith, Kiran V. Garimella, Mehrtash Babadi, Victoria Popic, Moshe Sade-Feldman, Michael Gatzen, et al. 2023. “High-Throughput RNA Isoform Sequencing Using Programmed CDNA Concatenation.” *Nature Biotechnology*, June. https://doi.org/10.1038/s41587-023-01815-7.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21 (1): 30.

Amsellem, Valerie, Nicola H. Dryden, Roberta Martinelli, Felicity Gavins, Lourdes Osuna Almagro, Graeme M. Birdsey, Dorian O. Haskard, Justin C. Mason, Patric Turowski, and Anna

M. Randi. 2014. “ICAM-2 Regulates Vascular Permeability and N-Cadherin Localization through Ezrin-Radixin-Moesin (ERM) Proteins and Rac-1 Signalling.” *Cell Communication and Signaling: CCS* 12 (1): 12.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10): R106.

Anvar, Seyed Yahya, Guy Allard, Elizabeth Tseng, Gloria M. Sheynkman, Eleonora de Klerk, Martijn Vermaat, Raymund H. Yin, et al. 2018. “Full-Length MRNA Sequencing Uncovers a Widespread Coupling between Transcription Initiation and MRNA Processing.” *Genome Biology*. https://doi.org/10.1186/s13059-018-1418-0.

Aragon, Jordon W., and Karen K. Hirschi. 2022. “Endothelial Cell Differentiation and Hemogenic Specification.” *Cold Spring Harbor Perspectives in Medicine*, February. https://doi.org/10.1101/cshperspect.a041164.

Ardui, Simon, Adam Ameur, Joris R. Vermeesch, and Matthew S. Hestand. 2018. “Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics.” *Nucleic Acids Research* 46 (5): 2159–68.

Arzalluz-Luque, Angeles, Pedro Salguero, Sonia Tarazona, and Ana Conesa. 2022. “Acorde Unravels Functionally Interpretable Networks of Isoform Co-Usage from Single Cell Data.” *Nature Communications* 13 (1): 1828.

Ascano, Manuel, Markus Hafner, Pavol Cekan, Stefanie Gerstberger, and Thomas Tuschl. 2012. “Identification of RNA-Protein Interaction Networks Using PAR-CLIP.” *Wiley Interdisciplinary Reviews. RNA* 3 (2): 159–77.

Atlasi, Yaser, Seyed J. Mowla, Seyed A. M. Ziaee, Paul J. Gokhale, and Peter W. Andrews. 2008. “OCT4 Spliced Variants Are Differentially Expressed in Human Pluripotent and Nonpluripotent Cells.” *Stem Cells*. https://doi.org/10.1634/stemcells.2008-0530.

Au, K. F., V. Sebastiano, P. T. Afshar, J. D. Durruthy, L. Lee, B. A. Williams, H. van Bakel, et al. 2013. “Characterization of the Human ESC Transcriptome by Hybrid Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (50): E4821-30.

Aydin, Selcan, Duy T. Pham, Tian Zhang, Gregory R. Keele, Daniel A. Skelly, Matthew Pankratz, Ted Choi, et al. 2022. “Genetic Dissection of the Pluripotent Proteome through Multi-Omics Data Integration.” *BioRxiv*. https://doi.org/10.1101/2022.04.22.489216.

Bainbridge, Matthew N., René L. Warren, Martin Hirst, Tammy Romanuik, Thomas Zeng, Anne Go, Allen Delaney, et al. 2006. “Analysis of the Prostate Cancer Cell Line LNCaP Transcriptome Using a Sequencing-by-Synthesis Approach.” *BMC Genomics* 7 (September): 246.

Banarjee, R., A. Sharma, S. Bai, A. Deshmukh, and M. Kulkarni. 2018. “Proteomic Study of Endothelial Dysfunction Induced by AGEs and Its Possible Role in Diabetic Cardiovascular Complications.” *Journal of Proteomics* 187: 69–79.

Baralle, F. E., and J. Giudice. 2017. “Alternative Splicing as a Regulator of Development and Tissue Identity.” *Nature Reviews. Molecular Cell Biology* 18 (7): 437–51.

Bargehr, Johannes, Lay Ping Ong, Maria Colzani, Hongorzul Davaapil, Peter Hofsteen, Shiv Bhandari, Laure Gambardella, et al. 2019. “Epicardial Cells Derived from Human Embryonic Stem Cells Augment Cardiomyocyte-Driven Heart Regeneration.” *Nature Biotechnology* 37 (8): 895–906.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. “Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry.” *Nature* 456 (7218): 53–59.

Berger, Michael F., and Martha L. Bulyk. 2006. “Protein Binding Microarrays (PBMs) for Rapid, High-Throughput Characterization of the Sequence Specificities of DNA Binding Proteins.” *Methods in Molecular Biology* 338: 245–60.

Berget, Susan M., Claire Moore, and Phillip A. Sharp. 1977. “Spliced Segments at the 5′ Terminus of Adenovirus 2 Late MRNA*.” *Proceedings of the National Academy of Sciences* 74 (8): 3171–75.

Bertero, Alessandro, Paul A. Fields, Vijay Ramani, Giancarlo Bonora, Galip G. Yardimci, Hans Reinecke, Lil Pabon, William S. Noble, Jay Shendure, and Charles E. Murry. 2019. “Dynamics of Genome Reorganization during Human Cardiogenesis Reveal an RBM20-Dependent Splicing Factory.” *Nature Communications* 10 (1): 1538.

Blakeley, Paul, Jennifer A. Siepen, Craig Lawless, and Simon J. Hubbard. 2010. “Investigating Protein Isoforms via Proteomics: A Feasibility Study.” *Proteomics* 10 (6): 1127–40.

Blanco, Francisco Javier, and Carmelo Bernabéu. 2012. “The Splicing Factor SRSF1 as a Marker for Endothelial Senescence.” *Frontiers in Physiology* 3 (March): 54.

Blech-Hermoni, Y., S. J. Stillwagon, and A. N. Ladd. 2013. “Diversity and Conservation of CELF1 and CELF2 RNA and Protein Expression Patterns during Embryonic Development.” *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 242 (6): 767–77.

Blencowe, Benjamin J. 2006. “Alternative Splicing: New Insights from Global Analyses.” *Cell* 126 (1): 37–47.

———. 2017. “The Relationship between Alternative Splicing and Proteomic Complexity.” *Trends in Biochemical Sciences* 42 (6): 407–8.

Bowler, E., and S. Oltean. 2019. “Alternative Splicing in Angiogenesis.” *International Journal of Molecular Sciences* 20 (9). https://doi.org/10.3390/ijms20092067.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 109 (January): 21.29.1-21.29.9.

Bulyk, Martha L., and A. J. Marian Walhout. 2013. “Gene Regulatory Networks.” In *Handbook of Systems Biology*, 65–88. Elsevier.

Cahan, P., H. Li, S. A. Morris, E. Lummertz da Rocha, G. Q. Daley, and J. J. Collins. 2014. “CellNet: Network Biology Applied to Stem Cell Engineering.” *Cell* 158 (4): 903–15.

Callaway, Ewen. 2024. “Chemistry Nobel Goes to Developers of AlphaFold AI That Predicts Protein Structures,” October. https://www.nature.com/articles/d41586-024-03214-7.

Caniuguir, A., B. J. Krause, C. Hernandez, R. Uauy, and P. Casanello. 2016. “Markers of Early Endothelial Dysfunction in Intrauterine Growth Restriction-Derived Human Umbilical Vein Endothelial Cells Revealed by 2D-DIGE and Mass Spectrometry Analyses.” *Placenta* 41 (May): 14–26.

Cao, Gaoyuan, Christopher D. O’Brien, Zhao Zhou, Samuel M. Sanders, Jordan N. Greenbaum, Antonis Makrigiannakis, and Horace M. DeLisser. 2002. “Involvement of Human PECAM-1 in Angiogenesis and in Vitro Endothelial Cell Migration.” *American Journal of Physiology. Cell Physiology* 282 (5): C1181-90.

Carlyle, B. C., R. R. Kitchen, J. Zhang, R. S. Wilson, T. T. Lam, J. S. Rozowsky, K. R. Williams, N. Sestan, M. B. Gerstein, and A. C. Nairn. 2018. “Isoform-Level Interpretation of High-Throughput Proteomics Data Enabled by Deep Integration with RNA-Seq.” *Journal of Proteome Research* 17 (10): 3431–44.

Castaldi, Peter J., Abdullah Abood, Charles R. Farber, and Gloria M. Sheynkman. 2022. “Bridging the Splicing Gap in Human Genetics with Long-Read RNA Sequencing: Finding the Protein Isoform Drivers of Disease.” *Human Molecular Genetics* 31 (R1): R123–36.

Castellana, N., and V. Bafna. 2010. “Proteogenomics to Discover the Full Coding Content of Genomes: A Computational Perspective.” *Journal of Proteomics* 73 (11): 2124–35.

Chang, Kuei-Ting, Lee-Hsin Wang, Yu-Mei Lin, Ching-Feng Cheng, and Guey-Shin Wang. 2021. “CELF1 Promotes Vascular Endothelial Growth Factor Degradation Resulting in Impaired Microvasculature in Heart Failure.” *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 35 (5): e21512.

Chen, Ying, Andre Sim, Yuk Kei Wan, Keith Yeo, Joseph Jing Xian Lee, Min Hao Ling, Michael I. Love, and Jonathan Göke. 2023. "Context-Aware Transcript Quantification from Long-Read RNA-Seq Data with Bambu." *Nature Methods* 20 (8): 1187–95.

Cheong, C. Y., and T. Lufkin. 2011. "Alternative Splicing in Self-Renewal of Embryonic Stem Cells." *Stem Cells International*. https://doi.org/10.4061/2011/560261.

Chepelev, Iouri, and Xin Chen. 2013. "Alternative Splicing Switching in Stem Cell Lineages." *Frontiers of Biology* 8 (1): 50–59.

Cleaver, O., and D. A. Melton. 2003. "Endothelial Signaling during Development." *Nature Medicine* 9 (6): 661–68.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17 (January): 13.

Cooper, Thomas A., Lili Wan, and Gideon Dreyfuss. 2009. "RNA and Disease." *Cell* 136 (4): 777–93.

Corey, D. R., and J. M. Abrams. 2001. "Morpholino Antisense Oligonucleotides: Tools for Investigating Vertebrate Development." *Genome Biology* 2 (5): REVIEWS1015.

Cox, David B. T., Jonathan S. Gootenberg, Omar O. Abudayyeh, Brian Franklin, Max J. Kellner, Julia Joung, and Feng Zhang. 2017. "RNA Editing with CRISPR-Cas13." *Science* 358 (6366): 1019–27.

Creighton, Harriet, and C. H. Waddington. 1958. "The Strategy of the Genes." *AIBS Bulletin* 8 (2): 49.

Cvitkovic, Ivan, and Melissa S. Jurica. 2013. "Spliceosome Database: A Tool for Tracking Components of the Spliceosome." *Nucleic Acids Research* 41 (Database issue): D132-41.

Dana, Hassan, Ghanbar Mahmoodi Chalbatani, Habibollah Mahmoodzadeh, Rezvan Karimloo, Omid Rezaiean, Amirreza Moradzadeh, Narges Mehmandoost, et al. 2017. "Molecular Mechanisms and Biological Functions of SiRNA." *International Journal of Biomedical Science: IJBS* 13 (2): 48–57.

Danan, Charles, Sudhir Manickavel, and Markus Hafner. 2016. "PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites." *Methods in Molecular Biology* 1358: 153–73.

Davis, R. L., H. Weintraub, and A. B. Lassar. 1987. "Expression of a Single Transfected CDNA Converts Fibroblasts to Myoblasts." *Cell* 51 (6): 987–1000.

Dejana, E., K. K. Hirschi, and M. Simons. 2017. "The Molecular Basis of Endothelial Cell Plasticity." *Nature Communications* 8: 14361.

Deslattes Mays, A., M. Schmidt, G. Graham, E. Tseng, P. Baybayan, R. Sebra, M. Sanda, J. B. Mazarati, A. Riegel, and A. Wellstein. 2019. “Single-Molecule Real-Time (SMRT) Full-Length RNA-Sequencing Reveals Novel and Distinct MRNA Isoforms in Human Bone Marrow Cell Subpopulations.” *Genes* 10 (4): 17.

Deutsch, Eric W., Christopher M. Overall, Jennifer E. Van Eyk, Mark S. Baker, Young-Ki Paik, Susan T. Weintraub, Lydie Lane, et al. 2016. “Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1.” *Journal of Proteome Research* 15 (11): 3961–70.

Di Matteo, Anna, Elisa Belloni, Davide Pradella, Ambra Cappelletto, Nina Volf, Serena Zacchigna, and Claudia Ghigna. 2020. “Alternative Splicing in Endothelial Cells: Novel Therapeutic Opportunities in Cancer Angiogenesis.” *Journal of Experimental & Clinical Cancer Research: CR* 39 (1): 275.

Dijk, E. L. van, Y. Jaszczyszyn, D. Naquin, and C. Thermes. 2018. “The Third Revolution in Sequencing Technology.” *Trends in Genetics: TIG* 34 (9): 666–81.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Dou, Zhihui, Dapeng Zhao, Xiaohua Chen, Caipeng Xu, Xiaodong Jin, Xuetian Zhang, Yupei Wang, et al. 2021. “Aberrant Bcl-x Splicing in Cancer: From Molecular Mechanism to Therapeutic Modulation.” *Journal of Experimental & Clinical Cancer Research: CR* 40 (1): 194.

Du, Menghan, Nathaniel Jillette, Jacqueline Jufen Zhu, Sheng Li, and Albert Wu Cheng. 2020. “CRISPR Artificial Splicing Factors.” *Nature Communications* 11 (1): 2973.

Dusserre, N., N. L’Heureux, K. S. Bell, H. Y. Stevens, J. Yeh, L. A. Otte, L. Loufrani, and J. A. Frangos. 2004. “PECAM-1 Interacts with Nitric Oxide Synthase in Human Endothelial Cells: Implication for Flow-Induced Nitric Oxide Synthase Activation.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 24 (10): 1796–1802.

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. 2009. “GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists.” *BMC Bioinformatics* 10 (February): 48.

Edwards, John M., Jed Long, Cornelia H. de Moor, Jonas Emsley, and Mark S. Searle. 2013. “Structural Insights into the Targeting of MRNA GU-Rich Elements by the Three RRMs of CELF1.” *Nucleic Acids Research* 41 (14): 7153–66.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. “Real-Time DNA Sequencing from Single Polymerase Molecules.” *Science* 323 (5910): 133–38.

Eisenstein, Michael. 2023. “Innovative Technologies Crowd the Short-Read Sequencing Market.” *Nature* 614 (7949): 798–800.

Elias, J. E., and S. P. Gygi. 2007. “Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry.” *Nature Methods* 4 (3): 207–14.

Endoh, Tamaki, and Takashi Ohtsuki. 2009. “Cellular SiRNA Delivery Using Cell-Penetrating Peptides Modified for Endosomal Escape.” *Advanced Drug Delivery Reviews* 61 (9): 704–9.

Erickson, Brian K., Christopher M. Rose, Craig R. Braun, Alison R. Erickson, Jeffrey Knott, Graeme C. McAlister, Martin Wühr, Joao A. Paulo, Robert A. Everley, and Steven P. Gygi. 2017. “A Strategy to Combine Sample Multiplexing with Targeted Proteomics Assays for High-Throughput Protein Signature Characterization.” *Molecular Cell* 65 (2): 361–70.

Fair, Benjamin, Carlos F. Buen Abad Najar, Junxing Zhao, Stephanie Lozano, Austin Reilly, Gabriela Mossian, Jonathan P. Staley, Jingxin Wang, and Yang I. Li. 2024. “Global Impact of Unproductive Splicing on Human Gene Expression.” *Nature Genetics* 56 (9): 1851–61.

Farnham, Peggy J. 2009. “Insights from Genomic Profiling of Transcription Factors.” *Nature Reviews. Genetics* 10 (9): 605–16.

Farrokh, S., A. L. Brillen, J. Haendeler, J. Altschmied, and H. Schaal. 2015. “Critical Regulators of Endothelial Cell Functions: For a Change Being Alternative.” *Antioxidants & Redox Signaling* 22 (14): 1212–29.

Ferrer-Bonsoms, Juan A., Ignacio Cassol, Pablo Fernández-Acín, Carlos Castilla, Fernando Carazo, and Angel Rubio. 2020. “ISOGO: Functional Annotation of Protein-Coding Splice Variants.” *Scientific Reports* 10 (1): 1069.

Fiszbein, A., and A. R. Kornblihtt. 2017. “Alternative Splicing Switches: Important Players in Cell Differentiation.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 39 (6). https://doi.org/10.1002/bies.201600157.

Floor, S. N., and J. A. Doudna. 2016. “Tunable Protein Synthesis by Transcript Isoforms in Human Cells.” *ELife* 5. https://doi.org/10.7554/eLife.10921.

Floyd, Brendan M., and Edward M. Marcotte. 2022. “Protein Sequencing, One Molecule at a Time.” *Annual Review of Biophysics* 51 (1): 181–200.

Frankish, A., M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.

Frankish, Adam, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E. Loveland, Jonathan M. Mudge, Cristina Sisu, et al. 2021. “GENCODE 2021.” *Nucleic Acids Research* 49 (D1): D916–23.

Franzén, Oscar, Li-Ming Gan, and Johan L. M. Björkegren. 2019. “PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data.” *Database: The Journal of Biological Databases and Curation* 2019 (January). https://doi.org/10.1093/database/baz046.

Fu, Xiang-Dong, and Manuel Ares Jr. 2014. “Context-Dependent Control of Alternative Splicing by RNA-Binding Proteins.” *Nature Reviews. Genetics* 15 (10): 689–701.

Gabut, Mathieu, Payman Samavarchi-Tehrani, Xinchen Wang, Valentina Slobodeniuc, Dave O’Hanlon, Hoon-Ki Sung, Manuel Alvarez, et al. 2011. “An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming.” *Cell* 147 (1): 132–46.

Gagliardi, Miriam, and Maria R. Matarazzo. 2016. “RIP: RNA Immunoprecipitation.” *Methods in Molecular Biology* 1480: 73–86.

Galas, D. J., and A. Schmitz. 1978. “DNAse Footprinting: A Simple Method for the Detection of Protein-DNA Binding Specificity.” *Nucleic Acids Research* 5 (9): 3157–70.

Gallien, Sebastien, Sang Yoon Kim, and Bruno Domon. 2015. “Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM)*[S].” *Molecular & Cellular Proteomics: MCP* 14 (6): 1630–44.

Gehman, Lauren T., Pratap Meera, Peter Stoilov, Lily Shiue, Janelle E. O’Brien, Miriam H. Meisler, Manuel Ares Jr, Thomas S. Otis, and Douglas L. Black. 2012. “The Splicing Regulator Rbfox2 Is Required for Both Cerebellar Development and Mature Motor Function.” *Genes & Development* 26 (5): 445–60.

Geuens, Thomas, Delphine Bouhy, and Vincent Timmerman. 2016. “The HnRNP Family: Insights into Their Role in Health and Disease.” *Human Genetics* 135 (8): 851–67.

Giampietro, C., G. Deflorian, S. Gallo, A. Di Matteo, D. Pradella, S. Bonomi, E. Belloni, et al. 2015. “The Alternative Splicing Factor Nova2 Regulates Vascular Development and Lumen Formation.” *Nature Communications* 6: 8479.

Godo, Shigeo, and Hiroaki Shimokawa. 2017. “Endothelial Functions.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 37 (9). https://doi.org/10.1161/atvbaha.117.309813.

Gong, Wuming, Satyabrata Das, Javier E. Sierra-Pagan, Erik Skie, Nikita Dsouza, Thijs A. Larson, Mary G. Garry, Edgar Luzete-Monteiro, Kenneth S. Zaret, and Daniel J. Garry. 2022. “ETV2 Functions as a Pioneer Factor to Regulate and Reprogram the Endothelial Lineage.” *Nature Cell Biology* 24 (5): 672–84.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews. Genetics* 17 (6): 333–51.

Gordon, Sean P., Elizabeth Tseng, Asaf Salamov, Jiwei Zhang, Xiandong Meng, Zhiying Zhao, Dongwan Kang, et al. 2015. “Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule MRNA Sequencing.” *PloS One* 10 (7). https://doi.org/10.1371/journal.pone.0132628.

GTEx Consortium. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580–85.

Guerra-Espinosa, Claudia, María Jiménez-Fernández, Francisco Sánchez-Madrid, and Juan M. Serrador. 2024. “ICAMs in Immunity, Intercellular Adhesion and Communication.” *Cells (Basel, Switzerland)* 13 (4): 339.

Gundry, Rebekah L., Paul W. Burridge, and Kenneth R. Boheler. 2011. “Pluripotent Stem Cell Heterogeneity and the Evolving Role of Proteomic Technologies in Stem Cell Biology.” *Proteomics* 11 (20): 3947–61.

Hafner, Markus, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. 2021. “CLIP and Complementary Methods.” *Nature Reviews Methods Primers* 1 (1): 1–23.

Han, H., U. Braunschweig, T. Gonatopoulos-Pournatzis, R. J. Weatheritt, C. L. Hirsch, K. C. H. Ha, E. Radovani, et al. 2017. “Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors.” *Molecular Cell* 65 (3): 539-+.

Hang, Xingyi, Peiyao Li, Zhifeng Li, Wubin Qu, Ying Yu, Hualing Li, Zhiyong Shen, et al. 2009. “Transcription and Splicing Regulation in Human Umbilical Vein Endothelial Cells under Hypoxic Stress Conditions by Exon Array.” *BMC Genomics* 10 (March): 126.

Hardwick, Simon A., Anoushka Joglekar, Paul Flicek, Adam Frankish, and Hagen U. Tilgner. 2019. “Getting the Entire Message: Progress in Isoform Sequencing.” *Frontiers in Genetics* 10 (August): 709.

He, Shiyang, Eugene Valkov, Sihem Cheloufi, and Jernej Murn. 2023. “The Nexus between RNA-Binding Proteins and Their Effectors.” *Nature Reviews. Genetics* 24 (5): 276–94.

Heather, James M., and Benjamin Chain. 2016. “The Sequence of Sequencers: The History of Sequencing DNA.” *Genomics* 107 (1): 1–8.

Hon, Ting, Kristin Mars, Greg Young, Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin, Nicholas Maurer, et al. 2020. “Highly Accurate Long-Read HiFi Sequencing Data for Five Complex Genomes.” *Scientific Data* 7 (1): 399.

Hood, Leroy, James R. Heath, Michael E. Phelps, and Biaoyang Lin. 2004. “Systems Biology and New Technologies Enable Predictive and Preventative Medicine.” *Science (New York, N.Y.)* 306 (5696): 640–43.

Hood, Leroy, and Lee Rowen. 2013. “The Human Genome Project: Big Science Transforms Biology and Medicine.” *Genome Medicine* 5 (9): 79.

Hoof, Dennis van, Jeroen Krijgsveld, and Christine Mummery. 2012. “Proteomic Analysis of Stem Cell Differentiation and Early Development.” *Cold Spring Harbor Perspectives in Biology* 4 (3). https://doi.org/10.1101/cshperspect.a008177.

Huang, Kie Kyon, Jiawen Huang, Jeanie Kar Leng Wu, Minghui Lee, Su Ting Tay, Vikrant Kumar, Kalpana Ramnarayanan, et al. 2021. “Long-Read Transcriptome Sequencing Reveals Abundant Promoter Diversity in Distinct Molecular Subtypes of Gastric Cancer.” *Genome Biology* 22 (1): 44.

Ibeh, Neke, Pradiptajati Kusuma, Chelzie Crenna Darusallam, Safarina Malik, Herawati Sudoyo, Davis J. McCarthy, and Irene Gallego Romero. 2024. “Profiling Genetically Driven Alternative Splicing across the Indonesian Archipelago.” *BioRxiv*. https://doi.org/10.1101/2024.05.07.593052.

Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. “The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community.” *Genome Biology* 17 (1): 239.

Jangi, Mohini, and Phillip A. Sharp. 2014. “Building Robust Transcriptomes with Master Splicing Factors.” *Cell* 159 (3): 487–98.

Jiang, Wei, and Liang Chen. 2021. “Alternative Splicing: Human Disease and Quantitative Analysis from High-Throughput Sequencing.” *Computational and Structural Biotechnology Journal* 19: 183–95.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions.” *Science* 316 (5830): 1497–1502.

Jolma, Arttu, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, et al. 2010. “Multiplexed Massively Parallel SELEX for Characterization of Human Transcription Factor Binding Specificities.” *Genome Research* 20 (6): 861–73.

Joung, Julia, Sai Ma, Tristan Tay, Kathryn R. Geiger-Schuller, Paul C. Kirchgatterer, Vanessa K. Verdine, Baolin Guo, et al. 2023. “A Transcription Factor Atlas of Directed Differentiation.” *Cell* 186 (1): 209-229.e26.

Katz, Yarden, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. 2010. “Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation.” *Nature Methods* 7 (12): 1009–15.

Kelemen, O., P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, and S. Stamm. 2013. “Function of Alternative Splicing.” *Gene* 514 (1): 1–30.

Kelly, Melissa A., and Karen K. Hirschi. 2009. “Signaling Hierarchy Regulating Human Endothelial Cell Development.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 29 (5): 718–24.

Kennedy, Marion, Geneve Awong, Christopher M. Sturgeon, Andrea Ditadi, Ross LaMotte-Mohs, Juan Carlos Zúñiga-Pflücker, and Gordon Keller. 2012. “T Lymphocyte Potential Marks the Emergence of Definitive Hematopoietic Progenitors in Human Pluripotent Stem Cell Differentiation Cultures.” *Cell Reports* 2 (6): 1722–35.

Keppetipola, Niroshika, Shalini Sharma, Qin Li, and Douglas L. Black. 2012. “Neuronal Regulation of Pre-MRNA Splicing by Polypyrimidine Tract Binding Proteins, PTBP1 and PTBP2.” *Critical Reviews in Biochemistry and Molecular Biology* 47 (4): 360–78.

Khan, Shawez, Federico Taverna, Katerina Rohlenova, Lucas Treps, Vincent Geldhof, Laura de Rooij, Liliana Sokol, et al. 2019. “EndoDB: A Database of Endothelial Cell Transcriptomics Data.” *Nucleic Acids Research* 47 (D1): D736–44.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. “HISAT: A Fast Spliced Aligner with Low Memory Requirements.” *Nature Methods* 12 (4): 357–60.

Kim, Jihoon, Bon-Kyoung Koo, and Juergen A. Knoblich. 2020. “Human Organoids: Model Systems for Human Biology and Medicine.” *Nature Reviews. Molecular Cell Biology* 21 (10): 571–84.

Kinney, Melissa A., Linda T. Vo, Jenna M. Frame, Jessica Barragan, Ashlee J. Conway, Shuai Li, Kwok-Kin Wong, et al. 2019. “A Systems Biology Pipeline Identifies Regulatory Networks for Stem Cell Engineering.” *Nature Biotechnology* 37 (7): 810–18.

Kofler, Natalie M., and Michael Simons. 2015. “Angiogenesis versus Arteriogenesis: Neuropilin 1 Modulation of VEGF Signaling.” *F1000prime Reports* 7 (March): 26.

Korchak, Jennifer A., Erin D. Jeffery, Saikat Bandyopadhyay, Ben T. Jordan, Micah D. Lehe, Emily F. Watts, Aidan Fenix, Mathias Wilhelm, and Gloria M. Sheynkman. 2024. “IS-PRM-Based Peptide Targeting Informed by Long-Read Sequencing for Alternative Proteome Detection.” *Journal of the American Society for Mass Spectrometry*, July. https://doi.org/10.1021/jasms.4c00119.

Kornblihtt, Alberto R., Ignacio E. Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J. Muñoz. 2013. “Alternative Splicing: A Pivotal Step between Eukaryotic Transcription and Translation.” *Nature Reviews. Molecular Cell Biology* 14 (3): 153–65.

Kosti, Idit, Predrag Radivojac, and Yael Mandel-Gutfreund. 2012. “An Integrated Regulatory Network Reveals Pervasive Cross-Regulation among Transcription and Splicing Factors.” *PLoS Computational Biology* 8 (7): e1002603.

Kovaka, Sam, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. 2019. “Transcriptome Assembly from Long-Read RNA-Seq Alignments with StringTie2.” *Genome Biology* 20 (1): 278.

Kukurba, Kimberly R., and Stephen B. Montgomery. 2015. “RNA Sequencing and Analysis.” *Cold Spring Harbor Protocols* 2015 (11): 951–69.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. “The Human Transcription Factors.” *Cell* 172 (4): 650–65.

Lanahan, Anthony, Xi Zhang, Alessandro Fantin, Zhen Zhuang, Felix Rivera-Molina, Katherine Speichinger, Claudia Prahst, et al. 2013. “The Neuropilin 1 Cytoplasmic Domain Is Required for VEGF-A-Dependent Arteriogenesis.” *Developmental Cell* 25 (2): 156–68.

Lau, Edward, Yu Han, Damon R. Williams, Cody T. Thomas, Rajani Shrestha, Joseph C. Wu, and Maggie P. Y. Lam. 2019. “Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome.” *Cell Reports* 29 (11): 3751-3765.e5.

Lee, Yeon, and Donald C. Rio. 2015. “Mechanisms and Regulation of Alternative Pre-MRNA Splicing.” *Annual Review of Biochemistry* 84 (March): 291–323.

Lesha, Emal, Haydy George, Mark M. Zaki, Cory J. Smith, Parastoo Khoshakhlagh, and Alex H. M. Ng. 2023. “A Survey of Transcription Factors in Cell Fate Control.” *Methods in Molecular Biology* 2594: 133–41.

Leung, Szi Kay, Aaron R. Jeffries, Isabel Castanho, Ben T. Jordan, Karen Moore, Jonathan P. Davies, Emma L. Dempster, et al. 2021. “Full-Length Transcript Sequencing of Human and Mouse Cerebral Cortex Identifies Widespread Isoform Diversity and Alternative Splicing.” *Cell Reports* 37 (7): 110022.

Li, Wei Vivian, Shan Li, Xin Tong, Ling Deng, Hubing Shi, and Jingyi Jessica Li. 2019. “AIDE: Annotation-Assisted Isoform Discovery with High Precision.” *Genome Research* 29 (12): 2056–72.

Li, Wenyuan, Chun-Chi Liu, Shuli Kang, Jian-Rong Li, Yu-Ting Tseng, and Xianghong Jasmine Zhou. 2016. “Pushing the Annotation of Cellular Activities to a Higher Resolution: Predicting Functions at the Isoform Level.” *Methods* 93 (January): 110–18.

Li, Y. I., D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, and J. K. Pritchard. 2018. “Annotation-Free Quantification of RNA Splicing Using LeafCutter.” *Nature Genetics* 50 (1): 151–58.

Li, Yapu, Ding Wang, Hongtao Wang, Xin Huang, Yuqi Wen, Bingrui Wang, Changlu Xu, et al. 2021. “A Splicing Factor Switch Controls Hematopoietic Lineage Specification of Pluripotent Stem Cells.” *EMBO Reports* 22 (1): e50535.

Liang, Ying, Haiyue Xu, Tao Cheng, Yujuan Fu, Hanwei Huang, Wenchang Qian, Junyan Wang, et al. 2022. “Gene Activation Guided by Nascent RNA-Bound Transcription Factors.” *Nature Communications* 13 (1): 7329.

Linares, Anthony J., Chia-Ho Lin, Andrey Damianov, Katrina L. Adams, Bennett G. Novitch, and Douglas L. Black. 2015. “The Splicing Regulator PTBP1 Controls the Activity of the Transcription Factor Pbx1 during Neuronal Differentiation.” *ELife* 4 (December): e09268.

Lindoso, Rafael Soares, Tais H. Kasai-Brunswick, Gustavo Monnerat Cahli, Federica Collino, Adriana Bastos Carvalho, Antonio Carlos Campos de Carvalho, and Adalberto Vieyra. 2019. “Proteomics in the World of Induced Pluripotent Stem Cells.” *Cells* 8 (7). https://doi.org/10.3390/cells8070703.

Liu, Y., A. Beyer, and R. Aebersold. 2016. “On the Dependency of Cellular Protein Levels on MRNA Abundance.” *Cell* 165 (3): 535–50.

Liu, Yansheng, Mar Gonzàlez-Porta, Sergio Santos, Alvis Brazma, John C. Marioni, Ruedi Aebersold, Ashok R. Venkitaraman, and Vihandha O. Wickramasinghe. 2017. “Impact of Alternative Splicing on the Human Proteome.” *Cell Reports* 20 (5): 1229–41.

López, A. J. 1995. “Developmental Role of Transcription Factor Isoforms Generated by Alternative Splicing.” *Developmental Biology* 172 (2): 396–411.

Lu, Yu, Yuin-Han Loh, Hu Li, Marcella Cesana, Scott B. Ficarro, Jignesh R. Parikh, Nathan Salomonis, et al. 2014. “Alternative Splicing of MBD2 Supports Self-Renewal in Human Pluripotent Stem Cells.” *Cell Stem Cell* 15 (1): 92–101.

Luco, Reini F., Qun Pan, Kaoru Tominaga, Benjamin J. Blencowe, Olivia M. Pereira-Smith, and Tom Misteli. 2010. “Regulation of Alternative Splicing by Histone Modifications.” *Science* 327 (5968): 996–1000.

MacArthur, Ben D., Avi Ma’ayan, and Ihor R. Lemischka. 2009. “Systems Biology of Stem Cell Fate and Cellular Reprogramming.” *Nature Reviews. Molecular Cell Biology* 10: 672.

Madugundu, A. K., C. H. Na, R. S. Nirujogi, S. Renuse, K. P. Kim, K. H. Burns, C. Wilks, et al. 2019. “Integrated Transcriptomic and Proteomic Analysis of Primary Human Umbilical Vein Endothelial Cells.” *Proteomics* 19 (15): e1800315.

Mahla, Ranjeet Singh. 2016. “Stem Cells Applications in Regenerative Medicine and Disease Therapeutics.” *International Journal of Cell Biology* 2016 (July): 6940283.

Mallick, Parag, Markus Schirle, Sharon S. Chen, Mark R. Flory, Hookeun Lee, Daniel Martin, Jeffrey Ranish, et al. 2007. “Computational Prediction of Proteotypic Peptides for Quantitative Proteomics.” *Nature Biotechnology* 25 (1): 125–31.

Mann, M., N. A. Kulak, N. Nagaraj, and J. Cox. 2013. “The Coming Age of Complete, Accurate, and Ubiquitous Proteomes.” *Molecular Cell* 49 (4): 583–90.

Mantere, Tuomo, Simone Kersten, and Alexander Hoischen. 2019. “Long-Read Sequencing Emerging in Medical Genetics.” *Frontiers in Genetics* 10 (May): 426.

Marasco, Luciano E., and Alberto R. Kornblihtt. 2022. “The Physiology of Alternative Splicing.” *Nature Reviews. Molecular Cell Biology*, October. https://doi.org/10.1038/s41580-022-00545-z.

Marcelo, K. L., L. C. Goldie, and K. K. Hirschi. 2013. “Regulation of Endothelial Cell Differentiation and Specification.” *Circulation Research* 112 (9): 1272–87.

Mardis, E. R. 2013. “Next-Generation Sequencing Platforms.” *Annual Review of Analytical Chemistry* 6: 287–303.

Marques-Coelho, Diego, Lukas da Cruz Carvalho Iohan, Ana Raquel Melo de Farias, Amandine Flaig, Brainbank Neuro–CEB Neuropathology Network, Jean-Charles Lambert, and Marcos Romualdo Costa. 2021. “Differential Transcript Usage Unravels Gene Expression Alterations in Alzheimer’s Disease Human Brains.” *Npj Aging and Mechanisms of Disease* 7 (1): 2.

Marziano, Corina, Gael Genet, and Karen K. Hirschi. 2021. “Vascular Endothelial Cell Specification in Health and Disease.” *Angiogenesis* 24 (2): 213–36.

Mattick, John S. 2018. “The State of Long Non-Coding RNA Biology.” *Non-Coding RNA* 4 (3). https://doi.org/10.3390/ncrna4030017.

Mazin, P. V., P. Khaitovich, M. Cardoso-Moreira, and H. Kaessmann. 2021. “Alternative Splicing during Mammalian Organ Development.” *Nature Genetics* 2021/05/05. https://doi.org/10.1038/s41588-021-00851-w.

McGarvey, P. B., A. Nightingale, J. Luo, H. Huang, M. J. Martin, C. Wu, and UniProt, Consortium. 2019. “UniProt Genomic Mapping for Deciphering Functional Effects of Missense Variants.” *Human Mutation* 40 (6): 694–705.

Mehlferber, Madison M., Erin D. Jeffery, Jamie Saquing, Ben T. Jordan, Leon Sheynkman, Mayank Murali, Gael Genet, Bipul R. Acharya, Karen K. Hirschi, and Gloria M. Sheynkman. 2022. “Characterization of Protein Isoform Diversity in Human Umbilical Vein Endothelial Cells via Long-Read Proteogenomics.” *RNA Biology* 19 (1): 1228–43.

Melé, M., P. G. Ferreira, F. Reverter, and D. S. DeLuca. 2015. “The Human Transcriptome across Tissues and Individuals.” https://science.sciencemag.org/content/348/6235/660.abstract?casa_token=Mgz9BI3vfzQAAAA:LLdOaKcjIEz7Tz-3O13tk7gt0FHiXgl-Y5AkKCnyqjIy6qNXtwm8v-C9K7QwWHWe20LPgKRY9CUA.

"Method of the Year 2022: Long-Read Sequencing." 2023. *Nature Methods* 20 (1): 1.

Miller, Rachel M., Ben T. Jordan, Madison M. Mehlferber, Erin D. Jeffery, Christina Chatzipantsiou, Simi Kaur, Robert J. Millikin, et al. 2022. "Enhanced Protein Isoform Characterization through Long-Read Proteogenomics." *Genome Biology* 23 (1): 1–28.

Mishra, Sambit K., Viraj Muthye, and Gaurav Kandoi. 2020. "Computational Methods for Predicting Functions at the MRNA Isoform Level." *International Journal of Molecular Sciences* 21 (16): 5686.

Molnár, A., and K. Georgopoulos. 1994. "The Ikaros Gene Encodes a Family of Functionally Diverse Zinc Finger DNA-Binding Proteins." *Molecular and Cellular Biology* 14 (12): 8292–8303.

Montañés-Agudo, Pablo, Simona Aufiero, Eva N. Schepers, Ingeborg van der Made, Lucia Cócera-Ortega, Auriane C. Ernault, Stéphane Richard, et al. 2023. "The RNA-Binding Protein QKI Governs a Muscle-Specific Alternative Splicing Program That Shapes the Contractile Function of Cardiomyocytes." *Cardiovascular Research*, January. https://doi.org/10.1093/cvr/cvad007.

Montes, Matías, Brianne L. Sanford, Daniel F. Comiskey, and Dawn S. Chandler. 2019. "RNA Splicing and Disease: Animal Models to Therapies." *Trends in Genetics: TIG* 35 (1): 68–87.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28.

Moulton, Jon D. 2007. "Using Morpholinos to Control Gene Expression." *Current Protocols in Nucleic Acid Chemistry / Edited by Serge L. Beaucage ... [et Al.]* Chapter 4 (1): Unit 4.30.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62.

Mthembu, Nonkululeko N., Zukile Mbita, Rodney Hull, and Zodwa Dlamini. 2017. "Abnormalities in Alternative Splicing of Angiogenesis-Related Genes and Their Role in HIV-Related Cancers." *HIV/AIDS - Research and Palliative Care* 9: 77–93.

Mudge, J. M., and J. Harrow. 2016. "The State of Play in Higher Eukaryote Gene Annotation." *Nature Reviews. Genetics* 17 (12): 758–72.

Naftelberg, Shiran, Ignacio E. Schor, Gil Ast, and Alberto R. Kornblihtt. 2015. "Regulation of Alternative Splicing through Coupling with Transcription and Chromatin Structure." *Annual Review of Biochemistry* 84: 165–98.

Nelson, Elizabeth A., Jingyao Qiu, Nicholas W. Chavkin, and Karen K. Hirschi. 2021. “Directed Differentiation of Hemogenic Endothelial Cells from Human Pluripotent Stem Cells.” *Journal of Visualized Experiments: JoVE*, no. 169 (March). https://doi.org/10.3791/62391.

Nesvizhskii, A. I., and R. Aebersold. 2005. “Interpretation of Shotgun Proteomic Data: The Protein Inference Problem.” *Molecular & Cellular Proteomics: MCP* 4 (10): 1419–40.

Nesvizhskii, Alexey I. 2014. “Proteogenomics: Concepts, Applications and Computational Strategies.” *Nature Methods* 11 (11): 1114–25.

Nesvizhskii, Alexey I., and Ruedi Aebersold. 2005. “Interpretation of Shotgun Proteomic Data.” *Molecular & Cellular Proteomics: MCP* 4 (10): 1419–40.

Ng, Alex H. M., Parastoo Khoshakhlagh, Jesus Eduardo Rojo Arias, Giovanni Pasquini, Kai Wang, Anka Swiersy, Seth L. Shipman, et al. 2021. “A Comprehensive Library of Human Transcription Factors for Cell Fate Engineering.” *Nature Biotechnology* 39 (4): 510–19.

Nichol, Donna, and Heidi Stuhlmann. 2012. “EGFL7: A Unique Angiogenic Signaling Factor in Vascular Development and Disease.” *Blood* 119 (6): 1345–52.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. “Expansion of the Eukaryotic Proteome by Alternative Splicing.” *Nature* 463 (7280): 457–63.

Niwa, Hitoshi. 2018. “The Principles That Govern Transcription Factor Network Functions in Stem Cells.” *Development* 145 (6). https://doi.org/10.1242/dev.157420.

Nordon, Ian, Ranjeet Brar, Robert Hinchliffe, Gillian Cockerill, Ian Loftus, and Matt Thompson. 2009. “The Role of Proteomic Research in Vascular Disease.” *Journal of Vascular Surgery* 49 (6): 1602–12.

Oh, Yujeong, and Jiwon Jang. 2019. “Directed Differentiation of Pluripotent Stem Cells by Transcription Factors.” *Molecules and Cells* 42 (3): 200–209.

Oksuz, Ozgur, Jonathan E. Henninger, Robert Warneford-Thomson, Ming M. Zheng, Hailey Erb, Kalon J. Overholt, Susana Wilson Hawken, et al. 2022. “Transcription Factors Interact with RNA to Regulate Genes.” *BioRxiv*. https://doi.org/10.1101/2022.09.27.509776.

PacBio. 2020. “Sequencing 101: The Evolution of DNA Sequencing Tools.” PacBio. March 25, 2020. https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature Genetics* 40 (12): 1413–15.

Panina, Yulia, Peter Karagiannis, Andreas Kurtz, Glyn N. Stacey, and Wataru Fujibuchi. 2020. “Human Cell Atlas and Cell-Type Authentication for Regenerative Medicine.” *Experimental & Molecular Medicine* 52 (9): 1443–51.

Pardo-Palacios, Francisco J., Angeles Arzalluz-Luque, Liudmyla Kondratova, Pedro Salguero, Jorge Mestre-Tomás, Rocío Amorín, Eva Estevan-Morió, et al. 2023. "SQANTI3: Curation of Long-Read Transcriptomes for Accurate Identification of Known and Novel Isoforms." *BioRxiv*, June. https://doi.org/10.1101/2023.05.17.541248.

Pardo-Palacios, Francisco J., Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, et al. 2024. "Systematic Assessment of Long-Read RNA-Seq Methods for Transcript Identification and Quantification." *Nature Methods* 21 (7): 1349–63.

Park, Sunyoung, Christine M. Sorenson, and Nader Sheibani. 2015. "PECAM-1 Isoforms, ENOS and Endoglin Axis in Regulation of Angiogenesis." *Clinical Science* 129 (3): 217–34.

Parrotta, Elvira Immacolata, Stefania Scalise, Domenico Taverna, Maria Teresa De Angelis, Gianmarco Sarro, Marco Gaspari, Gianluca Santamaria, and Giovanni Cuda. 2019. "Comprehensive Proteogenomic Analysis of Human Embryonic and Induced Pluripotent Stem Cells." *Journal of Cellular and Molecular Medicine* 23 (8): 5440–53.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19.

Patro, Rob, Stephen M. Mount, and Carl Kingsford. 2014. "Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms." *Nature Biotechnology* 32 (5): 462–64.

Paul, Lukas, Petra Kubala, Gudrun Horner, Michael Ante, Igor Holländer, Seitz Alexander, and Torsten Reda. 2016. "SIRVs: Spike-in RNA Variants as External Isoform Controls in RNA-Sequencing." *BioRxiv*. bioRxiv. https://doi.org/10.1101/080747.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nature Biotechnology* 33 (3): 290–95.

Phanstiel, Doug, Justin Brumbaugh, W. Travis Berggren, Kevin Conard, Xuezhu Feng, Mark E. Levenstein, Graeme C. McAlister, James A. Thomson, and Joshua J. Coon. 2008. "Mass Spectrometry Identifies and Quantifies 74 Unique Histone H4 Isoforms in Differentiating Human Embryonic Stem Cells." *Proceedings of the National Academy of Sciences of the United States of America* 105 (11): 4093–98.

Pozo, Fernando, Laura Martinez-Gomez, Thomas A. Walsh, José Manuel Rodriguez, Tomas Di Domenico, Federico Abascal, Jesús Vazquez, and Michael L. Tress. 2021. "Assessing the Functional Relevance of Splice Isoforms." *NAR Genomics and Bioinformatics* 3 (2): lqab044.

Pozo, Fernando, José Manuel Rodriguez, Laura Martínez Gómez, Jesús Vázquez, and Michael L. Tress. 2022. “APPRIS Principal Isoforms and MANE Select Transcripts Define Reference Splice Variants.” *Bioinformatics (Oxford, England)* 38 (Suppl_2): ii89–94.

Prinos, Panagiotis, Daniel Garneau, Jean-François Lucier, Daniel Gendron, Sonia Couture, Marianne Boivin, Jean-Philippe Brosseau, et al. 2011. “Alternative Splicing of SYK Regulates Mitosis and Cell Survival.” *Nature Structural & Molecular Biology* 18 (6): 673–79.

Privratsky, Jamie R., and Peter J. Newman. 2014. “PECAM-1: Regulator of Endothelial Junctional Integrity.” *Cell and Tissue Research* 355 (3): 607–19.

Qiu, J., S. Nordling, H. H. Vasavada, E. C. Butcher, and K. K. Hirschi. 2020. “Retinoic Acid Promotes Endothelial Cell Cycle Early G1 State to Enable Human Hemogenic Endothelial Cell Specification.” *Cell Reports* 33 (9): 108465.

Quattrone, Alessandro, and Erik Dassi. 2019. “The Architecture of the Human RNA-Binding Protein Regulatory Network.” *IScience* 21 (November): 706–19.

Rackham, O. J., J. Firas, H. Fang, M. E. Oates, M. L. Holmes, A. S. Knaupp, Consortium, Fantom, et al. 2016. “A Predictive Computational Framework for Direct Reprogramming between Human Cell Types.” *Nature Genetics* 48 (3): 331–35.

Rajendran, Peramaiyan, Thamaraiselvan Rengarajan, Jayakumar Thangavel, Yutaka Nishigaki, Dhanapal Sakthisekaran, Gautam Sethi, and Ikuo Nishigaki. 2013. “The Vascular Endothelium and Human Diseases.” *International Journal of Biological Sciences* 9 (10): 1057–69.

Rao, Sridhar, Shao Zhen, Sergei Roumiantsev, Lindsay T. McDonald, Guo-Cheng Yuan, and Stuart H. Orkin. 2010. “Differential Roles of Sall4 Isoforms in Embryonic Stem Cell Pluripotency.” *Molecular and Cellular Biology* 30 (22): 5364–80.

“Real-Time QRT-PCR.” n.d. Accessed December 2, 2022. https://www.ncbi.nlm.nih.gov/probe/docs/techqpcr/.

Reed, Brian D., Michael J. Meyer, Valentin Abramzon, Omer Ad, Omer Ad, Pat Adcock, Faisal R. Ahmad, et al. 2022. “Real-Time Dynamic Single-Molecule Protein Sequencing on an Integrated Semiconductor Device.” *Science (New York, N.Y.)* 378 (6616): 186–92.

Reixachs-Solé, Marina, and Eduardo Eyras. 2022. “Uncovering the Impacts of Alternative Splicing on the Proteome with Current Omics Techniques.” *Wiley Interdisciplinary Reviews. RNA* 13 (4): e1707.

Rhine, Christy L., Christopher Neil, Jing Wang, Samantha Maguire, Luke Buerer, Mitchell Salomon, Ijeoma C. Meremikwu, Juliana Kim, Natasha T. Strande, and William G. Fairbrother. 2022. “Massively Parallel Reporter Assays Discover de Novo Exonic Splicing Mutants in Paralogs of Autism Genes.” *PLoS Genetics* 18 (1): e1009884.

Rhoads, A., and K. F. Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89.

Richardson, M. R., X. Y. Lai, F. A. Witzmann, and M. C. Yoder. 2010. “Venous and Arterial Endothelial Proteomics: Mining for Markers and Mechanisms of Endothelial Diversity.” *Expert Review of Proteomics* 7 (6): 823–31.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.

Rockel, Sylvie, Marcel Geertz, and Sebastian J. Maerkl. 2012. “MITOMI: A Microfluidic Platform for in Vitro Characterization of Transcription Factor-DNA Interaction.” *Methods in Molecular Biology* 786: 97–114.

Rodriguez, Jose Manuel, Juan Rodriguez-Rivas, Tomás Di Domenico, Jesús Vázquez, Alfonso Valencia, and Michael L. Tress. 2018. “APPRIS 2017: Principal Isoforms for Multiple Gene Sets.” *Nucleic Acids Research* 46 (D1): D213–17.

Rowe, R. Grant, and George Q. Daley. 2019. “Induced Pluripotent Stem Cells in Disease Modelling and Drug Discovery.” *Nature Reviews. Genetics* 20 (7): 377–88.

Sabatier, Pierre, Christian M. Beusch, Amir A. Saei, Mike Aoun, Noah Moruzzi, Ana Coelho, Niels Leijten, et al. 2021. “An Integrative Proteomics Method Identifies a Regulator of Translation during Stem Cell Maintenance and Differentiation.” *Nature Communications* 12 (1): 6558.

Salomonis, Nathan, Phillip J. Dexheimer, Larsson Omberg, Robin Schroll, Stacy Bush, Jeffrey Huo, Lynn Schriml, et al. 2016. “Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium.” *Stem Cell Reports* 7 (1): 110–25.

Salovska, Barbora, Hongwen Zhu, Tejas Gandhi, Max Frank, Wenxue Li, George Rosenberger, Chongde Wu, et al. 2020. “Isoform-Resolved Correlation Analysis between MRNA Abundance Regulation and Protein Level Degradation.” *Molecular Systems Biology* 16 (3): e9170.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.

Sauteur, Loïc, Alice Krudewig, Lukas Herwig, Nikolaus Ehrenfeuchter, Anna Lenard, Markus Affolter, and Heinz-Georg Belting. 2014. “Cdh5/VE-Cadherin Promotes Endothelial Cell Interface Elongation via Cortical Actin Polymerization during Angiogenic Sprouting.” *Cell Reports* 9 (2): 504–13.

Schmidt, Mirko H. H., Frank Bicker, Iva Nikolic, Jeannette Meister, Tanja Babuke, Srdjan Picuric, Werner Müller-Esterl, Karl H. Plate, and Ivan Dikic. 2009. “Epidermal Growth Factor-

like Domain 7 (EGFL7) Modulates Notch Signalling and Affects Neural Stem Cell Renewal." *Nature Cell Biology* 11 (7): 873–80.

Scotti, Marina M., and Maurice S. Swanson. 2016. "RNA Mis-Splicing in Disease." *Nature Reviews. Genetics* 17 (1): 19–32.

Sharma, Arun, Samuel Sances, Michael J. Workman, and Clive N. Svendsen. 2020. "Multi-Lineage Human IPSC-Derived Platforms for Disease Modeling and Drug Discovery." *Cell Stem Cell* 26 (3): 309–29.

Sharon, Donald, Hagen Tilgner, Fabian Grubert, and Michael Snyder. 2013. "A Single-Molecule Long-Read Survey of the Human Transcriptome." *Nature Biotechnology* 31 (11): 1009–14.

Shaw, Timothy I., Bi Zhao, Yuxin Li, Hong Wang, Liang Wang, Brandon Manley, Paul A. Stewart, and Aleksandra Karolak. 2022. "Multi-Omics Approach to Identifying Isoform Variants as Therapeutic Targets in Cancer Patients." *Frontiers in Oncology* 12 (November): 1051487.

She, Yi-Min, Michael Rosu-Myles, Lisa Walrond, and Terry D. Cyr. 2012. "Quantification of Protein Isoforms in Mesenchymal Stem Cells by Reductive Dimethylation of Lysines in Intact Proteins." *Proteomics* 12 (3): 369–79.

Shen, Fei, Chenyang Hu, Xin Huang, Hao He, Deng Yang, Jirong Zhao, and Xiaozeng Yang. 2023. "Advances in Alternative Splicing Identification: Deep Learning and Pantranscriptome." *Frontiers in Plant Science* 14 (September): 1232466.

Shen, Shihao, Juw Won Park, Zhi-Xiang Lu, Lan Lin, Michael D. Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. 2014. "RMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data." *Proceedings of the National Academy of Sciences of the United States of America* 111 (51): E5593-601.

Sheynkman, Gloria M., Michael R. Shortreed, Anthony J. Cesnik, and Lloyd M. Smith. 2016. "Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation." *Annual Review of Analytical Chemistry* 9 (1): 521–45.

Shi, Yanhong, Haruhisa Inoue, Joseph C. Wu, and Shinya Yamanaka. 2017. "Induced Pluripotent Stem Cell Technology: A Decade of Progress." *Nature Reviews. Drug Discovery* 16 (2): 115–30.

Shishkova, Evgenia, Alexander S. Hebert, and Joshua J. Coon. 2016. "Now, More than Ever, Proteomics Needs Better Chromatography." *Cell Systems* 3 (4): 321–24.

Skene, Peter J., and Steven Henikoff. 2017. "An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites." *ELife* 6 (January). https://doi.org/10.7554/eLife.21856.

Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel. 2018. “Overview of Next-Generation Sequencing Technologies: Overview of next-Generation Sequencing.” *Et al [Current Protocols in Molecular Biology]* 122 (1): e59.

Smith, L. M., N. L. Kelleher, and Consortium for Top Down, Proteomics. 2013. “Proteoform: A Single Term Describing Protein Complexity.” *Nature Methods* 10 (3): 186–87.

Smith, Lloyd M., Jeffrey N. Agar, Julia Chamot-Rooke, Paul O. Danis, Ying Ge, Joseph A. Loo, Ljiljana Paša-Tolić, Yury O. Tsybin, Neil L. Kelleher, and Consortium for Top-Down Proteomics. 2021. “The Human Proteoform Project: Defining the Human Proteome.” *Science Advances* 7 (46): eabk0734.

Soemedi, Rachel, Kamil J. Cygan, Christy L. Rhine, Jing Wang, Charlston Bulacan, John Yang, Pinar Bayrak-Toydemir, Jamie McDonald, and William G. Fairbrother. 2017. “Pathogenic Variants That Alter Protein Code Often Disrupt Splicing.” *Nature Genetics* 49 (6): 848–55.

Solntsev, S. K., M. R. Shortreed, B. L. Frey, and L. M. Smith. 2018. “Enhanced Global Post-Translational Modification Discovery with MetaMorpheus.” *Journal of Proteome Research* 17 (5): 1844–51.

Soneson, Charlotte, Katarina L. Matthes, Malgorzata Nowicka, Charity W. Law, and Mark D. Robinson. 2016. “Isoform Prefiltering Improves Performance of Count-Based Methods for Analysis of Differential Transcript Usage.” *Genome Biology* 17 (January): 12.

Sriram, Gopu, Jia Yong Tan, Intekhab Islam, Abdul Jalil Rufaihah, and Tong Cao. 2015. “Efficient Differentiation of Human Embryonic Stem Cells to Arterial and Venous Endothelial Cells under Feeder- and Serum-Free Conditions.” *Stem Cell Research & Therapy* 6 (December): 261.

Stamm, S., S. Ben-Ari, I. Rafalska, Y. S. Tang, Z. Y. Zhang, D. Toiber, T. A. Thanaraj, and H. Soreq. 2005. “Function of Alternative Splicing.” *Gene* 344 (January): 1–20.

Steijger, T., J. F. Abril, P. G. Engstrom, F. Kokocinski, Rgasp Consortium, T. J. Hubbard, R. Guigo, J. Harrow, and P. Bertone. 2013. “Assessment of Transcript Reconstruction Methods for RNA-Seq.” *Nature Methods* 10 (12): 1177–84.

Sterne-Weiler, Timothy, Rocio Teresa Martinez-Nunez, Jonathan M. Howard, Ivan Cvitovik, Sol Katzman, Muhammad A. Tariq, Nader Pourmand, and Jeremy R. Sanford. 2013. “Frac-Seq Reveals Isoform-Specific Recruitment to Polyribosomes.” *Genome Research* 23 (10): 1615–23.

Stopfer, Lauren E., Aaron S. Gajadhar, Bhavin Patel, Sebastien Gallien, Dennie T. Frederick, Genevieve M. Boland, Ryan J. Sullivan, and Forest M. White. 2021. “Absolute Quantification of Tumor Antigens Using Embedded MHC-I Isotopologue Calibrants.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (37): e2111173118.

Sun, Hai-Jian, Zhi-Yuan Wu, Xiao-Wei Nie, and Jin-Song Bian. 2019. “Role of Endothelial Dysfunction in Cardiovascular Diseases: The Link between Inflammation and Hydrogen Sulfide.” *Frontiers in Pharmacology* 10: 1568.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell* 126 (4): 663–76.

———. 2016. “A Decade of Transcription Factor-Mediated Reprogramming to Pluripotency.” *Nature Reviews. Molecular Cell Biology* 17 (3): 183–93.

Tapial, J., K. C. H. Ha, T. Sterne-Weiler, A. Gohr, U. Braunschweig, A. Hermoso-Pulido, M. Quesnel-Vallieres, et al. 2017. “An Atlas of Alternative Splicing Profiles and Functional Associations Reveals New Regulatory Programs and Genes That Simultaneously Express Multiple Major Isoforms.” *Genome Research* 27 (10): 1759–68.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, Cecile Pereira, Francisco Jose Pardo-Palacios, Hector del Risco, Marc Ferrell, et al. 2017. “SQANTI: Extensive Characterization of Long Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification.” *BioRxiv*, March. https://doi.org/10.1101/118083.

Tatetsu, H., N. R. Kong, G. Chong, G. Amabile, D. G. Tenen, and L. Chai. 2016. “SALL4, the Missing Link between Stem Cells, Development and Cancer.” *Gene* 584 (2): 111–19.

Thomas, James D., Jacob T. Polaski, Qing Feng, Emma J. De Neef, Emma R. Hoppe, Maria V. McSharry, Joseph Pangallo, et al. 2020. “RNA Isoform Screens Uncover the Essentiality and Tumor-Suppressor Activity of Ultraconserved Poison Exons.” *Nature Genetics* 52 (1): 84–94.

Thomson, J. A., J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. 1998. “Embryonic Stem Cell Lines Derived from Human Blastocysts.” *Science* 282 (5391): 1145–47.

Tilgner, Hagen, David G. Knowles, Rory Johnson, Carrie A. Davis, Sudipto Chakrabortty, Sarah Djebali, João Curado, Michael Snyder, Thomas R. Gingeras, and Roderic Guigó. 2012. “Deep Sequencing of Subcellular RNA Fractions Shows Splicing to Be Predominantly Co-Transcriptional in the Human Genome but Inefficient for LncRNAs.” *Genome Research* 22 (9): 1616–25.

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. “TopHat: Discovering Splice Junctions with RNA-Seq.” *Bioinformatics (Oxford, England)* 25 (9): 1105–11.

Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–15.

Tress, Michael L., Federico Abascal, and Alfonso Valencia. 2017a. “Alternative Splicing May Not Be the Key to Proteome Complexity.” *Trends in Biochemical Sciences* 42 (2): 98–110.

———. 2017b. “Most Alternative Isoforms Are Not Functionally Important.” *Trends in Biochemical Sciences*.

Trincado, Juan L., Juan C. Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J. Elliott, and Eduardo Eyras. 2018. “SUPPA2: Fast, Accurate, and Uncertainty-Aware Differential Splicing Analysis across Multiple Conditions.” *Genome Biology* 19 (1): 40.

Ule, Jernej, and Benjamin J. Blencowe. 2019. “Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution.” *Molecular Cell* 76 (2): 329–45.

UniProt, Consortium. 2019. “UniProt: A Worldwide Hub of Protein Knowledge.” *Nucleic Acids Research* 47 (D1): D506–15.

Van Hoof, Dennis, Albert J. R. Heck, Jeroen Krijgsveld, and Christine L. Mummery. 2008. “Proteomics and Human Embryonic Stem Cells.” *Stem Cell Research* 1 (3): 169–82.

Van Nostrand, Eric L., Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, et al. 2020. “A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins.” *Nature* 583 (7818): 711–19.

Van Nostrand, Eric L., Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, et al. 2016. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (ECLIP).” *Nature Methods* 13 (6): 508–14.

Venables, J. P., L. Lapasset, G. Gadea, P. Fort, R. Klinck, M. Irimia, E. Vignal, et al. 2013. “MBNL1 and RBFOX2 Cooperate to Establish a Splicing Programme Involved in Pluripotent Stem Cell Differentiation.” *Nature Communications* 4: 2480.

Verbruggen, S., S. Gessulat, R. Gabriels, A. Matsaroki, H. Van de Voorde, B. Kuster, S. Degroeve, et al. 2021. “Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics.” *Molecular & Cellular Proteomics: MCP* 2021/04/07: 100076.

Veschini, Lorenzo, Heba Sailem, Disha Malani, Vilja Pietiäinen, Ana Stojiljkovic, Erika Wiseman, and Davide Danovi. 2021. “High-Content Imaging to Phenotype Human Primary and IPSC-Derived Cells.” *Methods in Molecular Biology* 2185: 423–45.

Vierbuchen, Thomas, and Marius Wernig. 2012. “Molecular Roadblocks for Cellular Reprogramming.” *Molecular Cell* 47 (6): 827–38.

Vitting-Seerup, Kristoffer, and Albin Sandelin. 2019. “IsoformSwitchAnalyzeR: Analysis of Changes in Genome-Wide Patterns of Alternative Splicing and Its Functional Consequences.” *Bioinformatics* 35 (21): 4469–71.

Vogel, C., and E. M. Marcotte. 2012. “Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses.” *Nature Reviews. Genetics* 13 (4): 227–32.

Wachtel, Chaim, and James L. Manley. 2009. “Splicing of MRNA Precursors: The Role of RNAs and Proteins in Catalysis.” *Molecular BioSystems* 5 (4): 311–16.

Wang, D., B. Eraslan, T. Wieland, B. Hallstrom, T. Hopf, D. P. Zolg, J. Zecha, et al. 2019. “A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues.” *Molecular Systems Biology* 15 (2): e8503.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. 2008. “Alternative Isoform Regulation in Human Tissue Transcriptomes.” *Nature* 456 (7221): 470–76.

Wang, Jianlong, Jennifer J. Trowbridge, Sridhar Rao, and Stuart H. Orkin. 2008. *Proteomic Studies of Stem Cells*. Harvard Stem Cell Institute.

Wang, L., H. J. Park, S. Dasari, S. Wang, J. P. Kocher, and W. Li. 2013. “CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model.” *Nucleic Acids Research* 41 (6): e74.

Wang, Robert, Ingo Helbig, Andrew C. Edmondson, Lan Lin, and Yi Xing. 2023. “Splicing Defects in Rare Diseases: Transcriptomics and Machine Learning Strategies towards Genetic Diagnosis.” *Briefings in Bioinformatics* 24 (5): bbad284.

Wang, X., and J. W. Dai. 2010. “Concise Review: Isoforms of OCT4 Contribute to the Confusing Diversity in Stem Cell Biology.” *Stem Cells* 28 (5): 885–93.

Wang, X., R. J. Slebos, D. Wang, P. J. Halvey, D. L. Tabb, D. C. Liebler, and B. Zhang. 2012. “Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data.” *Journal of Proteome Research* 11 (2): 1009–17.

Wang, Xiaojing, Simona G. Codreanu, Bo Wen, Kai Li, Matthew C. Chambers, Daniel C. Liebler, and Bing Zhang. 2018. “Detection of Proteome Diversity Resulted from Alternative Splicing Is Limited by Trypsin Cleavage Specificity.” *Molecular & Cellular Proteomics: MCP* 17 (3): 422–30.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews. Genetics* 10 (1): 57–63.

Weirick, Tyler, Giuseppe Militello, Raphael Müller, David John, Stefanie Dimmeler, and Shizuka Uchida. 2016. “The Identification and Characterization of Novel Transcripts from RNA-Seq Data.” *Briefings in Bioinformatics* 17 (4): 678–85.

Wells, C. A., and J. Choi. 2019. “Transcriptional Profiling of Stem Cells: Moving from Descriptive to Predictive Paradigms.” *Stem Cell Reports* 13 (2): 237–46.

White, Amanda Louise, and Gregory Jaye Bix. 2023. “VEGFA Isoforms as Pro-Angiogenic Therapeutics for Cerebrovascular Diseases.” *Biomolecules* 13 (4). https://doi.org/10.3390/biom13040702.

Wichmann, Christoph, Florian Meier, Sebastian Virreira Winter, Andreas-David Brunner, Jürgen Cox, and Matthias Mann. 2019. “MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides.” *Molecular & Cellular Proteomics: MCP* 18 (5): 982a–9994.

Wichmann, Christoph, Florian Meier, Sebastian Virreira Winter, Andreas-David Brunner, Jürgen Cox, and Matthias Mann. n.d. “MaxQuant.Live Enables Global Targeting of More than 25,000 Peptides.” https://doi.org/10.1101/443838.

Wilkinson, A. C., H. Nakauchi, and B. Gottgens. 2017. “Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity.” *Cell Systems* 5 (4): 319–31.

Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018.

Wineberg, Yishay, Itamar Kanter, Nissim Ben-Haim, Naomi Pode-Shakked, Efrat Bucris, Tali Hana Bar-Lev, Sarit Oriel, et al. 2022. “Characterization of Alternative MRNA Splicing in Cultured Cell Populations Representing Progressive Stages of Human Fetal Kidney Development.” *Scientific Reports* 12 (1): 19548.

Wiśniewski, Jacek R. 2018. “Filter-Aided Sample Preparation for Proteome Analysis.” *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-8695-8_1.

Wright, Jane E., and Rafal Ciosk. 2013. “RNA-Based Regulation of Pluripotency.” *Trends in Genetics: TIG* 29 (2): 99–107.

Xue, Yuanchao, Kunfu Ouyang, Jie Huang, Yu Zhou, Hong Ouyang, Hairi Li, Gang Wang, et al. 2013. “Direct Conversion of Fibroblasts to Neurons by Reprogramming PTB-Regulated MicroRNA Circuits.” *Cell* 152 (1–2): 82–96.

Yamanaka, S. 2008. “Induction of Pluripotent Stem Cells from Mouse Fibroblasts by Four Transcription Factors.” *Cell Proliferation* 41 Suppl 1 (Suppl 1): 51–56.

Yang, Xinping, Jasmin Coulombe-Huntington, Shuli Kang, Gloria M. Sheynkman, Tong Hao, Aaron Richardson, Song Sun, et al. 2016. “Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing.” *Cell* 164 (4): 805–17.

Yeo, Jia-Chi, and Huck-Hui Ng. 2013. “The Transcriptional Regulation of Pluripotency.” *Cell Research* 23 (1): 20–32.

Young, Richard A. 2011. “Control of the Embryonic Stem Cell State.” *Cell* 144 (6): 940–54.

Yu, Qing, Haopeng Xiao, Mark P. Jedrychowski, Devin K. Schweppe, Jose Navarrete-Perea, Jeffrey Knott, John Rogers, Edward T. Chouchani, and Steven P. Gygi. 2020. “Sample Multiplexing for Targeted Pathway Proteomics in Aging Mice.” *Proceedings of the National Academy of Sciences* 117 (18): 9723–32.

Zaro, Balyn W., Joseph J. Noh, Victoria L. Mascetti, Janos Demeter, Benson George, Monika Zukowska, Gunsagar S. Gulati, et al. 2020. “Proteomic Analysis of Young and Old Mouse Hematopoietic Stem Cells and Their Progenitors Reveals Post-Transcriptional Regulation in Stem Cells.” *ELife* 9 (November). https://doi.org/10.7554/eLife.62210.

Zhang, Tao, Yu Lin, Jing Liu, Zi Guan Zhang, Wei Fu, Li Yan Guo, Lei Pan, et al. 2016. “Rbm24 Regulates Alternative Splicing Switch in Embryonic Stem Cell Cardiac Lineage Differentiation.” *Stem Cells* 34 (7): 1776–89.

Zhou, Qiao, Juliana Brown, Andrew Kanarek, Jayaraj Rajagopal, and Douglas A. Melton. 2008. “In Vivo Reprogramming of Adult Pancreatic Exocrine Cells to Beta-Cells.” *Nature* 455 (7213): 627–32.

Zhu, Lili, Krishna Choudhary, Barbara Gonzalez-Teran, Yen-Sin Ang, Reuben Thomas, Nicole R. Stone, Lei Liu, et al. 2022. “Transcription Factor GATA4 Regulates Cell Type-Specific Splicing Through Direct Interaction With RNA in Human Induced Pluripotent Stem Cell-Derived Cardiac Progenitors.” *Circulation* 146 (10): 770–87.