UNIVERSITY OF CALIFORNIA

SANTA CRUZ

BUILDING A BETTER TRANSCRIPTOME

A dissertation submitted in partial satisfaction of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

In

MOLECULAR, CELLULAR, DEVELOPMENTAL BIOLOGY

By

Ashley L. Byrne

December 2019

The dissertation of Ashley L. Byrne is approved:

Professor Chris Vollmers, Chair

Professor Camilla Forsberg

Professor Susan Carpenter

Quentin Williams
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **iv**

List of Tables **vii**

Abstract **viii**

Dedication **ix**

Acknowledgments **x**

Introduction **1**

1. Tn5Prime, a Tn5 based 5' Capture Method for Single Cell RNA-seq **30**
2. Nanopore Single Cell RNAseq Reveals Widespread Transcriptional Variation Among Surface Receptors of B cells **73**
3. Depletion of Hemoglobin Transcripts and Long Read Sequencing Improves Annotation of Polar Bear (Ursus maritimus) **115**
4. Conclusion **158**

Bibliography **164**

List of Figures

0.1 **Fundamental difference between short- and long-read sequencing of transcripts** 10

0.2 **Long-read transcriptome sequencing approaches don't cover long transcripts** 14

0.3 **Error-prone reads pose analysis challenge** 20

0.4 **R2C2 method overview** 21

0.5 **Increase R2C2 subreads decreases indels and mismatches** 22

0.6 **Analysis challenges of long-read full-length sequencing** 24

1.1 **Tn5Prime Library construction and 5' capture** 37

1.2 **Tn5Prime peaks are highly concordant with GENCODE annotation and CAGE peaks** 40

1.3 **Tn5Prime quantifies transcriptomes accurately and reproducibly** 42

1.4 **Transcription start sites are detected in single CD27^{high} CD38^{high} B cells** 47

1.5 **Clustering of CD27^{high} CD38^{high} B cells** 49

1.6 **Assembling Antibody transcripts from Tn5Prime data** 53

S1.1 **Tn5Prime detects B cell specific genes in single B2 cells** 65

S1.2 **FACS Gating strategy for the isolation of CD27^{high} CD38^{high} human B cells** 66

S1.3 **Quality Metrics Based on Alignment Percentage and Read number** 67

S1.4 **tSNE on subsampled data** 68

S1.5 **Clustering of CD27+ CD38+ B cells** 69

S1.6 **Heavy Chain Pairings Determined by Read Depth** 70

2.1 **Experimental Design and ONT read characteristics** 80

2.2 **ONT RNAseq recapitulates Illumina RNAseq gene expression quantification** 82

2.3 **Quantifying gene/transcript expression with ONT RNAseq data** 84

2.4 **Identifying and quantifying isoforms in SIRV E2 mixture** 89

2.5 **ONT RNAseq data identifies isoform features in mouse B1A cells** 91

2.6 **Uncovering isoform diversity in B cell surface receptors** 95

S2.1 **ONT Single Cell Sequencing run characteristics** 108

S2.2 **Gene detection with subsampled data** 109

S2.3 **Quantifying SIRV transcripts amplified from 10fg starting material with ONT RNAseq** 110

S2.4 **Identifying SIRV Transcript isoforms using ONT RNAseq** 111

S2.5 **Diverse Isoforms of B cell surface receptors identified using ONT RNAseq** 112

3.1 **Schematic of Long-DASH** 122

3.2 **Long-DASH depletes hemoglobin from full-length cDNA** 123

3.3 **Long-DASH specifically targets hemoglobin from cDNA** 126

3.4 **Long-DASH depletes hemoglobin transcripts from cDNA** 128

3.5 **R2C2/Mandalorion identifies new isoform features in the polar bear transcriptome** 132

3.6 **R2C2/Mandalorion refine transcriptome annotations** 133

S3.1 **Long-DASH sgRNA depletes human hemoglobin transcripts from full-length cDNA** 147

S3.2 **Depletion of hemoglobin affects expression levels within but not fold-change between samples** 148

S3.3 **Alignment of orthologous HBA and HBB mRNA sequences in human and polar bear** 149

S3.4 **sgRNA design and construction** 153

4.1 **Simple schematic of Phi29 amplification** 162

4.2 **PCR-free R2C2 method improves transcript read length** 163

List of Tables

0.1 **Sequencing technology characteristics** 7

S1.1 **Oligos used in the Tn5 Prime manuscript** 71

S2.1 **List of Oligos** 113

S2.2 **Illumina and ONT Read Numbers** 114

S2.3 **Alternative Splice Site Predictions** 114

S3.1 **High-throughput sequencing runs and characteristics** 156

S3.2 **Oligos used in the Long-DASH** 157

Abstract

Building A Better Transcriptome

By Ashley L. Byrne

From a single embryo to billions of cells, a whole organism is constructed in a carefully regulated symphony. Every cell in our body shares the same sheet of music in the form of DNA but it is through RNA transcription that rhyme and meter are kept; comprised of a complex regulatory system determining when genes get turned on or off. It is through this regulation that alterations occur, allowing two identical cells to ultimately give rise to completely separate organs and tissue systems. Our current understanding of these processes relies heavily on using short-read sequencing technology to analyze whole transcriptomes. However, this method requires fragmentation of full-length molecules, making it difficult to recapitulate the transcriptome landscape in its entirety. Requiring heavy computational tools to assemble the transcriptome, which only provides an estimation. This loss of contiguity makes it clear we cannot depend on short-read RNA sequencing alone to truly understand the complexities within our transcriptome. Thus, I have established a toolset for creating better, more precise transcriptomes from single cells to bulk RNA studies. This body of work entails how we can elucidate transcript features that tend to be lost in short-read sequencing data. These improvements include developing a 5' capturing method for single cell data, employing a long-read single cell full-length cDNA sequencing method, increasing throughput and limiting length bias for bulk transcriptomic studies. Together, these improvements create a better snapshot of the transcriptome and will help change how we analyze transcriptomes.

Dedication

To my sweet beloved husband, Gabriel Byrne

Who is the only person I would ever trust during a Zombie Apocalypse

I love you very much

And to all the rebel girls

who want to rule the world.

Acknowledgments

I decided to write my acknowledgments on the last day in my lab before heading off to my next adventure, which may have been a bad idea. First and foremost, I would like to thank my mentor, Dr. Chris Vollmers. His immense love of science is inspirational and many might believe that people who go into science are doing it for glory or have an ego but we are just zealots in search of the unknown and yearn to solve problems. His scientific fervor has led me to believe that not only is science amazing but it is also fun. He was such an excellent mentor and I consider him a good friend.

My mother, Kristine Levitoff, who is one of the toughest woman I know. She has definitely taught me perseverance and grit. Science can sometimes be difficult, unpredictable and cruel but we have to move forward even if the results are good or bad.

My father, Brad Brooks, who was always supportive of my education and like myself had dreams that were once thought to be unattainable but has persevered and is doing what he loves. He is truly inspirational.

I would also like to acknowledge my lab mate, Charles Cole who is an amazing Bioinformatician. I look forward to seeing what amazing things he does next. I will miss not only our scientific discussions but our political discussions as well. I am a strong advocate of putting a Biologist and Bioinformatician next to each other because you can learn so much from one another.

Over the course of my graduate career there are many individuals who have given me advice, guidance and most importantly provided fun diversions to keep me occupied.

I finally, would like to acknowledge my husband, Gabriel Byrne. When we first met, I had no clue where our lives would lead. I knew that we both had ambitions to learn more and wanted to sharpen our senses of the world. I am so lucky to have a partner where we can discuss our scientific interests and “totally nerd out” as my mother puts it.

Realizing The Potential of Full-Length Transcriptome Sequencing

[This section is adapted from a review, **Realizing the Potential of Full-length Transcriptome Sequencing**, accepted from Philosophical Transactions B]

Ashley Byrne[1,*], Charles Cole[2,*], Roger Volden[2,*], Christopher Vollmers[2,#]

1) Department of Molecular, Cellular, and Developmental Biology, University of California Santa Cruz, CA 95064

2) Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064

* Contributed equally to this manuscript.
Names in alphabetical order.

# Correspondence should be addressed to:
Dr. Christopher Vollmers
vollmers@ucsc.edu

Abstract

Long-read sequencing holds great potential for transcriptome analysis because it offers researchers an affordable method to annotate transcriptomes for not only well-studied organisms but for less researched organisms such as non-models. However, non-model organisms have much more to gain using this technology as they cannot rely on large consortia projects to generate these transcriptome annotations. To utilize this technology to its full potential, several remaining molecular and computational challenges will have to be overcome. In this review, we have outlined the limitations of short-read sequencing technology and how long-read sequencing technology overcomes these limitations. We have also highlighted the unique challenges still present for long-read sequencing technology and provided some suggestions on how to overcome these challenges going forward.

Introduction

The rapid progress and application of sequencing technology after the completion of the Human Genome Project has led to a vastly expanded knowledge of the genome sequences present in the eukaryotic tree of life. However, due to cost and technological limitations, truly high-quality genome references have been limited to a core of organisms of large scientific or economic interest. Further, our knowledge of which parts of genomes constitute genes and which transcript isoforms these genes produce, i.e. high-quality transcriptome annotation, is even more scarce (Salzberg 2019). However, sequencing technology might be reaching a point where it will become feasible to affordably generate high-quality genome references and transcriptome annotations of a much wider range of organisms.

High-throughput sequencing technologies have grown massively more powerful over the last decade. During this time the ability to assemble genomes has outpaced the ability to annotate the transcriptomes they produce. Genome assembly is now entering a golden age where, for a moderate investment, high-quality "centromere-to-telomere" genome sequences can be assembled through a mix of several technologies, including short-read sequencing, linked short-read sequencing (HiC), long-read sequencing and optical mapping (Jain et al. 2018; Putnam et al. 2016; Dixon et al. 2018). These powerful and relatively affordable approaches are going to be of outsize benefit for non-model organisms from unicellular eukaryotes to polar bears that in the past did not receive the attention and large sums of money required to generate a high-quality genome reference the hard way - Chromosome maps, Sanger sequencing of BAC libraries, etc.

However, while we are quickly reaching a point at which genomes can relatively reliably be assembled into chromosome-scale scaffolds, transcriptome annotation lags behind in its ability to identify the genes and isoforms expressed from these chromosomes.

Transcriptome annotations are required for us to understand how genome sequences and changes to these sequences are interpreted by the cellular machinery. They are also required for many functional analyses. The process of genome annotation, using RNAseq often relies heavily on machine learning using in-silico ab algorithms to predict protein coding genes. However, these predictions become less accurate when dealing with organism specific protein coding genes when intron-exon boundaries and transcript features such as transcription start (TSS) and end sites (TES) differ substantially. Thus, without accurate transcriptome annotations, it can be difficult to investigate differential expression of mRNA isoforms or predict which proteins are present in a particular tissue or organism. Further, transcriptome annotation allows us to modify cellular behavior by allowing the design of siRNA or gRNA sequences that will effectively silence the expression of a targeted gene. This can especially be problematic for researchers who cannot accurately identify these features causing headaches for those trying to perform knock-in or knock-down experiments.

Currently, short- and long-read sequencing are used for transcriptome annotation but have underlying limitations that make reaching a “reference-level” transcriptome annotation both highly labor intensive and often simply impossible.

Here we discuss the potential and limitations of long-read based full-length transcriptome sequencing for transcriptome annotation and lay out a path towards realizing said potential.

1) What are the limitations of short-read sequencing technology?

The analysis of what RNA transcripts (annotation) are present in a sample and at what level (quantification) has relied on a mix of technologies over the last three decades. Early efforts to annotate and quantify complex eukaryotic transcriptomes were highly labor intensive. During the early 1990’s, efforts to evaluate RNA sequences on a large scale relied heavily on ESTs (Expressed Sequence Tags) whereby cDNA molecules were individually cloned, screened, and Sanger-sequenced to determine full-length mRNA sequences and observe semi-quantitative changes in gene expression (Adams et al. 1992). The Sanger-sequencing based SAGE (Serial Analysis of Gene Expression) method improved quantification and reduced cost by concatenating smaller 15-20 bp fragments of many cDNA molecules together for sequencing (Velculescu et al. 1995). However, because of the short length of analyzed fragments SAGE was inherently less useful for annotation. Hybridization-based microarray approaches completely eschewed annotation but simplified the quantification of already annotated genes (Lockhart et al. 1996).

The introduction of massively parallel sequencing in the mid-to-late 2000s completely changed transcriptome annotation and quantification. When massively parallel sequencing – best represented by the now dominant Illumina technology – became available to research labs it could generate millions of sequencing reads at a length of ~30 nucleotides (nt). Although initially intended for the sequencing of genomic DNA, researchers quickly found ways to leverage the power of these sequencers for transcriptome analysis in the form of the RNA-seq assay. RNA-seq sequences short cDNA fragments at extremely high throughput and quickly displaced microarray-based transcriptome analysis for a number of reasons including cost considerations as well as the ability to detect previously unknown transcripts and quantify the use of individual splice sites. In the last decade, Illumina sequencers have steadily and massively improved, although these improvements have come with compromises in experimental design. Most prominently newer Illumina sequencers require additional precautions to avoid sample cross-contamination during the sequencing reaction (Sinha et al. 2017).

Current Illumina sequencers like the NovaSeq can generate billions of sequencing reads at a length of 250 nt allowing the multiplexed analysis of hundreds to thousands of samples in a single run (Table 0.1). At this read-length and output, RNA-seq reads aren't only useful for transcriptome quantification but also for annotation. Consequently, efforts like GENCODE and RefSeq heavily rely on this data type for their respective annotation approaches (Harrow et al. 2012; Pruitt et al. 2014). Paired with literally hundreds of sample preparation techniques and analysis pipelines, transcriptome analysis by short-read RNA-seq (Mortazavi et al. 2008) is now a core component of research in nearly all fields of biology.

So, while it is clear that RNA-seq has revolutionized transcriptome annotation and quantification it is also becoming increasingly clear that it is ultimately a stop-gap solution of limited power born out the limitations of short-read sequencing. These limitations prevent RNA-seq from annotating and quantifying transcriptomes on the level of RNA transcript isoforms, i.e. transcript variants expressed by the same gene utilizing combinations of alternative splice sites, transcription start sites, and transcription termination or polyA sites. Thus, to fully understand the fundamentals of gene expression, isoform information is crucial.

Technology	**Read Throughput per $1K**	**Accuracy**	**Base Accuracy**	**Max Read Length**
Illumina NextSeq	~2x10^8	99.9%	N/A	75-300 bp
Pacific Biosciences (PacBio) Sequel	~4x10^5	89%	>99%	50,000 bp
Oxford Nanopore Technologies (ONT) MinION	~5x10^6	88%	>97.5%*	Up to 2Mb

Table 0.1: Sequencing technology characteristics. Read number per dollar is hard to establish considering different pricing structures and instrument costs. Here, we assume a lab would use sequencing cores for Illumina and PacBio sequencing while performing Oxford Nanopore Technologies (ONT) MinION sequencing themselves. *Consensus accuracy using our R2C2 approach as published (Volden et al., 2018).

a) Limitations in transcriptome assembly algorithms

Despite its dominant position in transcriptome analysis, short-read RNA-seq has failed at capturing the true complexity of eukaryotic transcriptomes. While RNA-seq can interrogate individual transcript features like splice sites, transcription start sites, and polyA sites, it fails at determining how these individual features are combined together into comprehensive transcript isoforms. This is due to the fact that the read length of short-read sequencers is too short to capture entire transcripts from end-to-end (Fig. 0.1). Incomplete fragments of transcripts therefore have to be computationally assembled into full-length isoforms. This is done using powerful algorithms performing de-novo (e.g. Trinity, rnaSPAdes (Grabherr et al. 2011; Bankevich et al. 2012)) or genome-guided transcriptome assemblies (e.g. Cufflinks, StringTie (Pertea et al. 2015; Trapnell et al. 2010)). All of these assemblers ultimately fail at discerning complex transcript isoforms expressed by the same gene because of limitations of the underlying data. First, RNA-seq reads often do not cover the ends of transcripts leaving TSS and polyA sites unresolved (Picelli, Faridani, et al. 2014). Second, although early studies have indicated that the mean exon size in the human genome is 147 bp, large exons > 300 bp are thought to occur in about 5% of the genome (Lander et al. 2001; Bolisetty and Beemon 2012). Large exons that are > 300bp may be too far apart to be resolved by normal short-read RNA-seq raw data, i.e. if a transcript has two alternative splice sites 1000 bp apart, no individual RNA-seq read will ever connect those two events. Given the relatively small exons sizes, exon chaining whereby exons are linked together by identifying all splice junctions is still problematic due to unequal read coverage. Computational methods that take this into account have been developed, however they still fail at deconvoluting complex isoform mixtures (Kanitz et al. 2015).

b) Limitations in short-read sequencing technology

To date, no RNA-seq protocol has succeeded in providing data capable of overcoming this assembly challenge and recovering full-length isoforms in a high-throughput manner. Although short-read sequencing technology has increased its sequencing length capability to ~300 nt from the original ~30 nt, it still can’t sequence the vast majority of transcripts from end-to-end. To get around the read length limitation, creative specialized protocols have been developed. The most successful methods include Synthetic Long Read (SLR) and spISO-seq which operates on the principle of splitting one sample into hundreds or thousands of separate reactions using either 384-well plates or microdroplets (Tilgner et al. 2015, 2017). This separation allows the generation of individual sequencing libraries that ideally only contain one transcript isoform for any specific gene. These libraries can then be sequenced and analyzed separately which massively simplifies computational assembly and reduces mis assemblies. However, while improving on general RNA-seq methods, neither method succeeds at effectively generating transcript isoforms. SLR generates a low number of transcripts most of which are incomplete at the 3’ end while spISO-seq generates sparse “read-clouds” that can connect individual splice sites but fail at consistently capturing 5’ and 3’ ends of transcripts. Additionally, both SLR and spISO-seq approaches have complex library preparation workflows that cannot be multiplexed and require specialized instrumentation which has prevented them from being widely adopted.

While it is not inconceivable that a future short-read based protocol will ultimately succeed at isoform-level analysis, this task currently appears to be well beyond the capabilities of short-read sequencing.

2) How can the potential of long-read transcriptome sequencing be realized?

We believe that long-read sequencing is on the verge of transforming transcriptome analysis similarly to how short-read sequencing did a decade ago. In contrast to short-read sequencing, long-read sequencing technology as provided by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) has the potential to identify and quantify isoforms simply by sequencing cDNA or mRNA molecules end-to-end from 3' polyA tail to 5' CAP.

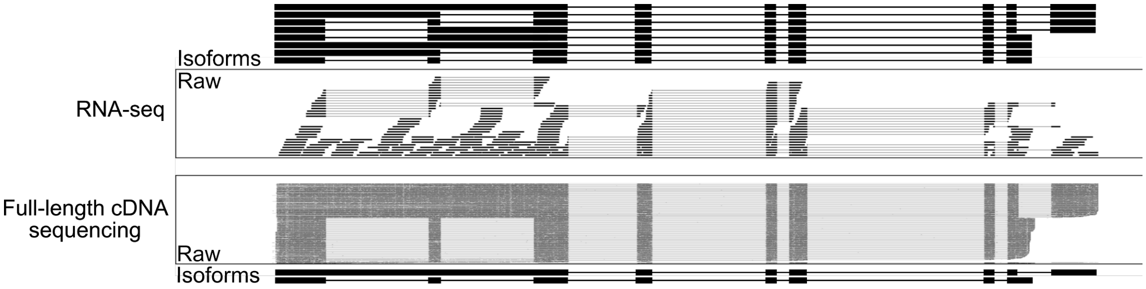


Figure 0.1: Fundamental difference between short- and long-read sequencing of transcripts. Short RNA-seq reads only capture small fragments of transcripts. RNA-seq data therefore lacks unambiguous isoform data leading to the inference of many erroneous isoforms. Long-read full-length cDNA data captures transcripts end-to-end making isoform inference unambiguous.

Just like short-read sequencing, long-read technology was initially intended for genomic DNA sequencing, but it was only a matter of time until cDNA copies of RNA transcript molecules were sequenced on PacBio and ONT sequencers.

Initial studies used long reads for the targeted analysis of specific highly complex transcripts (Treutlein, Gokce, et al. 2014) or to add small amounts of long-read data to short-read RNA-seq data (Koren et al. 2012; Au et al. 2013). Increasing read throughput has allowed the analysis of whole transcriptomes of diverse organisms with long-read data alone (B. Wang et al. 2016; Sharon et al. 2013; Tilgner et al. 2013, 2014) and in addition to the analysis of cDNA, ONT sequencers now offer the ability to sequence RNA directly (Workman et al. 2018; Garalde et al. 2018). Finally, long-read technology has been used to analyze the transcriptomes of single cells (Gupta et al. 2018; Volden et al. 2018; Byrne et al. 2017).

These papers clearly highlight the potential of long-read sequencing to identify new isoforms and isoform features like new splice sites, TSSs, and polyA sites which is essential to unambiguously annotate and quantify transcriptomes. These papers also lay out a path for the future: In the short-term, long-read technology will be a boon for the transcriptome annotation. With a moderate investment generating long-read transcriptome data for a variety of tissues and organs present in a non-model organism, transcriptome annotations will get closer to the comprehensiveness and quality of highly curated mouse and human transcriptomes. In the long-term, we believe long-read technology has the potential to entirely replace short-read RNA-seq for transcriptome analysis. However, to realize this potential, long-read transcriptome analysis still has to overcome several challenges that are currently limiting its progress.

3) What are the challenges of long-read sequencing?

Although the above examples have highlighted the potential of long-read technology, there still remains significant challenges which affect both PacBio and ONT to varying degrees: a) RNA integrity, b) length bias, c) read throughput, d) read accuracy, and 5) data analysis.

In order for long-read sequencing to be a main driver in pushing the transcriptome field forward these challenges will have to be overcome:

a) RNA integrity

All current long-read transcriptome sequencing approaches suffer from experimental artifacts caused by degraded RNA molecules. While ONT and PacBio sequencers make it possible to sequence entire transcripts from end-to-end, this only matters if the vast majority of sequenced transcript molecules are fully intact. The integrity of RNA going into long-read sequencing experiments is therefore of the highest importance. However, it is not yet clear what represents the best extraction and processing method for RNA.

Single-cell studies circumvent this issue by performing reverse transcription (RT) directly on cell lysates resulting in high quality results (Byrne et al. 2017; Gupta et al. 2018), but this is not the case for bulk samples comprised of tissues or many cells because highly concentrated cell lysates tend to inhibit RT reactions. Current efforts to dissociate, lyse, and extract RNA from bulk samples mostly rely on physical disruption and Trizol or Tri-reagent based protocols. These protocols are either followed by precipitations often resulting in Phenol and Guanidium contamination which can compromise RNA integrity or require a column-based clean-up potentially fragmenting long RNA transcripts in a way similar to high molecular weight genomic DNA.

Going forward we will need systematic studies comparing extraction methods for the integrity for very long transcripts (>10kb) which cannot be measured by the frequently used RIN value which is calculated by evaluating the integrity of the much shorter rRNA transcripts at ~2kb (18S) and ~5kb (28S). Additionally, similar to how Spike-In RNA Variants (SIRVs) or ERCC spike-ins are used to validate quantification it is possible that 5' capped synthetic transcripts could help indicate the integrity of RNA transcripts to identify the percentage of RNA degradation occurring within a given method (Cronin et al. 2004; Hardwick et al. 2016).

We believe these efforts are likely to succeed. Moving from short-read to long-read sequencing has already led to the genomics community rethinking the way it extracts DNA - from mostly column-based to precipitation-based approaches - leading to the successful sequencing of DNA molecules almost 1 million base-pairs in length (Jain et al. 2018).

b) Length bias

All current long-read transcriptome sequencing approaches are biased towards shorter transcripts. As a result, the read lengths produced do not reflect the transcript lengths as determined by annotation efforts like GENCODE. While the expression of short and long transcripts surely varies for each sample and each sample will only include a fraction of all transcripts in the GENCODE annotation, the fact remains that current long-read approaches appear to have a hard time capturing long transcripts.

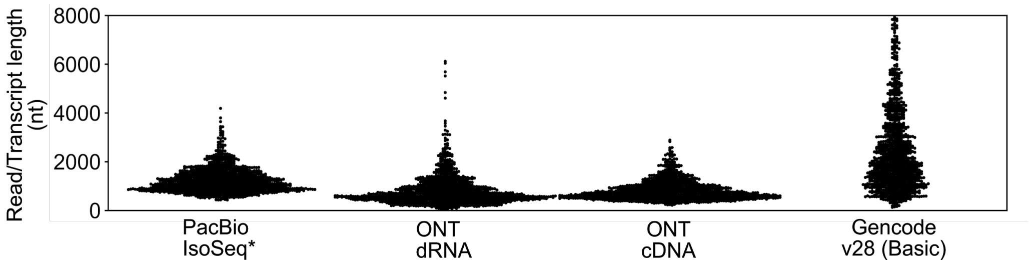


Figure 0.2: Long-read transcriptome sequencing approaches don't cover long transcripts. Swarmplots of length distributions of 1000 randomly sampled PacBio (Tilgner et al. 2014), ONT dRNA, and cDNA (Workman et al. 2018) reads covering the GM12878 (human lymphoblast cell line) transcriptome. These distributions are not representative of the length distribution of the human transcriptome as annotated by GENCODE. *While we show the most recent data set on GM12878 we could find for PacBio technology it is several years old and might not be fully representative of current platform performance.

This bias is rooted in the way samples are prepared for sequencing as well as the sequencing technology itself. To prepare full-length eukaryotic mRNA molecules for sequencing, protocols for PacBio and ONT sequencers today rely on some version of reverse transcription (RT) using oligo-dT priming most often paired with template switching as featured in the Smart-seq2 protocol (Picelli, Faridani, et al. 2014). This reverse transcription step generates cDNA with known 3' and 5' ends that can be PCR amplified. PCR amplification is required to generate enough cDNA for sequencing library preparation - several micrograms for either technology. However, if cDNA is PCR amplified, shorter transcripts are more likely to be successfully amplified, thereby generating a pool of cDNA skewed towards full-length short transcripts (<2kb) and shorter amplification artifacts of long transcripts.

While ONT sequencers can now sequence RNA directly, recent studies have shown that this does not overcome RNA degradation or length-bias issues. In fact, incomplete transcript sequences represent the majority of the produced data and this issue gets exacerbated with increased transcript length making direct RNA sequencing currently challenging for transcripts over 2kb in length (Fig. 0.2) (Workman et al. 2018).

In addition to length biases of sample preparation, both PacBio and ONT sequencers themselves have a bias towards shorter molecules. To systematically test this bias derived from different RNA extraction, sample preparation and sequencing methods new approaches will be needed. Unfortunately, current synthetic RNA spike-in mixtures like ERCC and SIRV (Lexogen), only contain molecules <2.5kb which is simply not long enough to determine bias against long transcript molecules (Fig. 0.2). To truly determine length bias, it would require sequencing of a well-defined eukaryotic human transcriptome, e.g. human cell line GM12878, using an RNA molecule length independent short-read RNA-seq method. While short-read RNA-seq won't be able to systematically resolve isoforms, assemblies of these reads can be used to estimate transcript lengths. Comparing representation of transcripts of different lengths in long-read data sets prepared with different protocols will then help reveal biases of these protocols.

The question then still remains: How do we overcome the inherent limitations of PCR amplification, sequencing library preparation, and the cDNA and direct RNA sequencing process itself. One thing is for certain future efforts will have to overcome these limitations or, ironically, the world of long transcripts will remain closed to long-read transcriptomics.

It will be up to the wider genomics community as well as PacBio and ONT to address these limitations. While adding complexity to sample preparations and distorting sample compositions, size selections on the RNA or cDNA level might mitigate length bias in sample preparation. Also, reducing cDNA amounts required for sequencing reactions might eliminate the need for PCR entirely. Additionally, PacBio sequencers have already made big strides reducing the length bias of their library preparation and sequencing reactions in the last few years and it would be surprising if this wasn't a big focus of ONT as well. Finally, one way to get around the length-bias of sequencing library preparation and sequencing reactions themselves is to dissociate transcript length from the length of the DNA/RNA being sequenced, i.e. making all DNA/RNA going into a sequencing reaction approximately the same length. This could be done by randomly ligating transcripts into large chimeric molecules or generating large DNA concatemers containing many copies of the same transcript (Volden et al. 2018).

c) Read throughput

All sequencing based transcriptome analysis is ultimately limited by the number of reads available for analysis. More reads result in better data but so far there hasn't been a rigorous study to identify the exact numbers of long sequencing reads required to exhaustively analyze a complex eukaryotic transcriptome. Because sequencing-based transcriptome analysis follows the same sampling principle regardless of read length, it stands to reason that these numbers will be similar to those required for short-read RNA-seq assays. Therefore, >30 million reads will be required for a shallow analysis of a transcriptome of a bulk sample capturing the isoforms of genes with medium and high expression (Sims et al. 2014). This however represents an ideal scenario assuming a single isoform per gene. If we think about treating individual isoforms as individual genes, it follows that significantly deeper sequencing will be needed to identify and quantify them. Indeed, it has been suggested by a deep-sequencing survey of alternative-splicing in human tissue that there are, on average, seven alternative splicing events per multi-exon gene (Pan et al. 2008). Therefore, to truly explore the complexity of mammalian transcriptomes, >100,000,000 of reads covering full-length transcripts will be required per tissue or organ.

In contrast to bulk samples, estimating the read depth required for single cell analysis is more straightforward as it is limited by experimental constraints. Most workflows in the rapidly expanding field of single cell transcriptome analysis attach unique molecular identifiers (UMIs) to each cDNA molecule generated for each individual cell, thereby giving us a direct way to determine the number of reads needed to capture all or most of these molecules (Ziegenhain et al. 2017). 10X Genomics single cell analysis approach for example generates <20,000 cDNA molecules per cell (Zheng et al. 2017). To reliably capture >90% of these molecules, sampling statistics dictate the need for ~45,000 sequencing reads per cell and consequently 45,000,000 sequencing reads for the analysis of a 1,000 cell cDNA pool.

In short, future long-read transcriptome analysis of bulk and single cell samples will require tens to hundreds of millions of reads at a reasonable cost. While ONT sequencers now routinely generate several millions of reads per $1000 of sequencing, PacBio sequencers produce ~400,000 reads per $1000 of sequencing (Table 0.1). This means that achieving the sequencing depth required for exhaustive transcriptome analysis is now borderline feasible with ONT sequencers but would deplete all but the largest of research budgets if using PacBio sequencers. It will be interesting to see how the newly released ONT PromethION and PacBio Sequel II will change this equation once in researchers' hands as they both represent significant improvements in read throughput over the ONT MinION and PacBio Sequel, respectively.

d) Read accuracy

As long as a PacBio or ONT read captures the sequence of a full-length transcript and is accurate enough to be correctly aligned to a single genomic location, it is useful for analysis. There is no line in the metaphorical read-accuracy sand beyond which this transcript sequence becomes useless for analysis, because different downstream applications will require different levels of accuracy to be implemented. It is, however, no surprise that more accurate reads are always preferable over less accurate reads.

Both PacBio and ONT long-read technologies sequence individual DNA (or RNA) molecules and as such are inherently more error-prone than short-read Illumina sequencing which can rely on the combined signal of thousands of copies of DNA molecules to determine base sequence. Because the raw read length of PacBio sequencers is much longer than an average transcript molecule, circularized cDNA molecules can be read multiple times to generate a more accurate consensus. As a result, PacBio's IsoSeq protocol generates cDNA circular consensus sequences (CCS) that can achieve >99% (Q20) accuracy (Table 0.1) (Gupta et al. 2018; Tilgner et al. 2014).

While ONT raw read length far exceeds transcript length, there currently exists no commercial product to - like PacBio's CCS approach - take advantage of this read length to improve read accuracy through consensus generation. Because of this, cDNA or direct RNA sequencing on ONT (1D) generates sequences of 88% (Q9) accuracy (Fig. 0.3).

This low accuracy creates some serious drawbacks regarding downstream analysis including the inability to accurately demultiplex single cell data. Single cell approaches like 10X Genomics Chromium workflow or the Drop-seq protocol can process many hundreds to thousands of cells in parallel using water-in-oil emulsions to produce highly multiplexed single cell cDNA pools (Zheng et al. 2017; Macosko et al. 2015). In this process, cell-specific identifiers - short nucleotide sequences - are attached to each cDNA molecule that is reverse transcribed from mRNA. Consequently, assigning a cDNA molecule to the cell it originated from, i.e. demultiplexing, requires accurately determining the sequence of its cell-specific identifier. Without sufficiently accurate sequencing, molecules will therefore be mis-assigned or lost (Gupta et al. 2018).

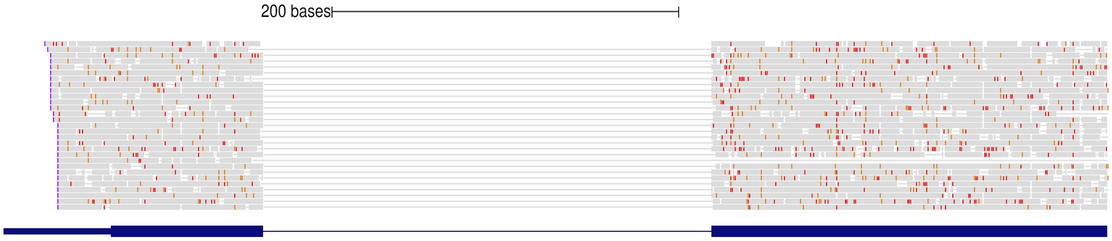


Figure 0.3: Error-prone reads pose analysis challenge. Representative alignments of ONT cDNA (Workman et al. 2018) reads. 30 read alignments (grey) to the first two exons of the CD19 gene (dark blue) are shown. Read alignments contain many insertions (orange), mismatches (red), and deletions (thin line) within exons. These errors complicate detection of exact transcript sequences and exact positions of splice sites, TSSs, and polyA sites.

Currently, ONT is working on improving their basecalling accuracy and have announced a commercial consensus approach to be released in 2019 that should address this issue. Until then the ONT research community including our own laboratory has recognized this issue and developed consensus sequencing approaches (C. Li et al. 2016). Specifically targeted for cDNA, the R2C2 approach we developed circularizes cDNA and uses rolling circle amplification to generate long concatemeric molecules that can be sequenced and processed into consensus sequences (Volden et al. 2018) (Fig. 0.4). At $1000 sequencing cost the R2C2 approach can currently produce several million sequencing reads at >97.5% (~Q16) (Fig 0.5) median accuracy (Byrne et al. 2019). Additionally, the R2C2 approach has proven very useful for de-multiplexing single cell data where 74% of the total reads containing cellular indexes derived from 7nt and 8nt combinations were able to be assigned. Given ONT's new basecalling algorithm we are confident this number will eventually increase.

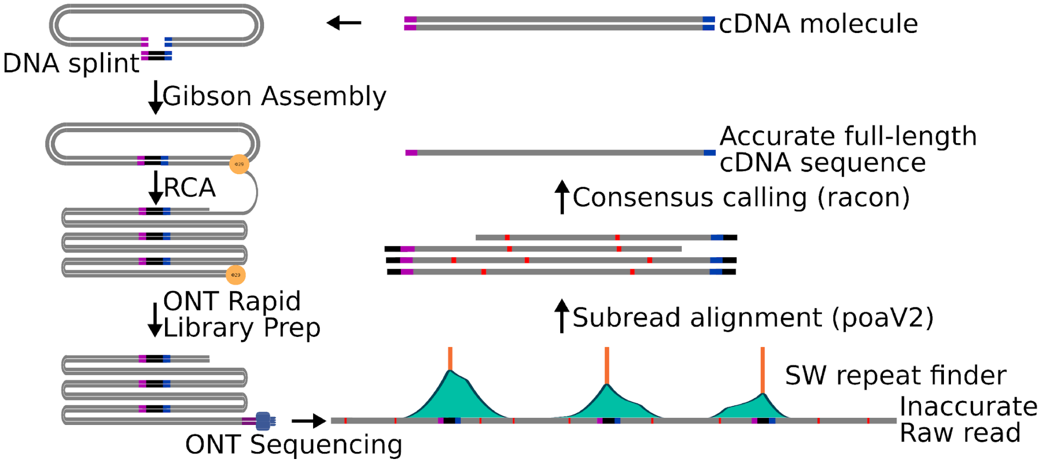


Figure 0.4: R2C2 method overview. Figure adapted from (Volden et al., 2018). cDNA is circularized using Gibson Assembly, amplified using RCA, and sequenced using the ONT MinION. The resulting raw reads are split into subreads containing full-length or partial cDNA sequences, which are combined into an accurate consensus sequence using our C3POa workflow, which relies on a custom algorithm to detect DNA splints as well as poaV2 and racon.

It is unclear whether ONT consensus approaches will be able to reach the accuracy of PacBio circular consensus reads, since the errors in PacBio sequencing data aren't entirely random they are less systematic than ONT errors (Weirather et al. 2017). Systematic errors which recur in the same base context, e.g. around homopolymers (stretches of the same base longer than 5nt) can pose insurmountable challenges for consensus-based error-correction. Error-correcting algorithms like *Nanopolish* (Loman, Quick, and Simpson 2015), *Racon* (Vaser et al. 2017), or *Medaka* (Medaka n.d.) are beginning to address this by either making use of the ionic current based raw signal generated by ONT sequencers or by incorporating ONT specific error models. While it is not yet clear what accuracy will be sufficient for reliably identifying regular transcript isoforms, increasing the accuracy of individual reads to beyond 99% will not only be required for single cell cDNA demultiplexing but also the analysis of individual transcripts that contain unique sequences not encoded in the genome, i.e. B and T cell receptor transcripts, as well as transcripts containing base modifications.

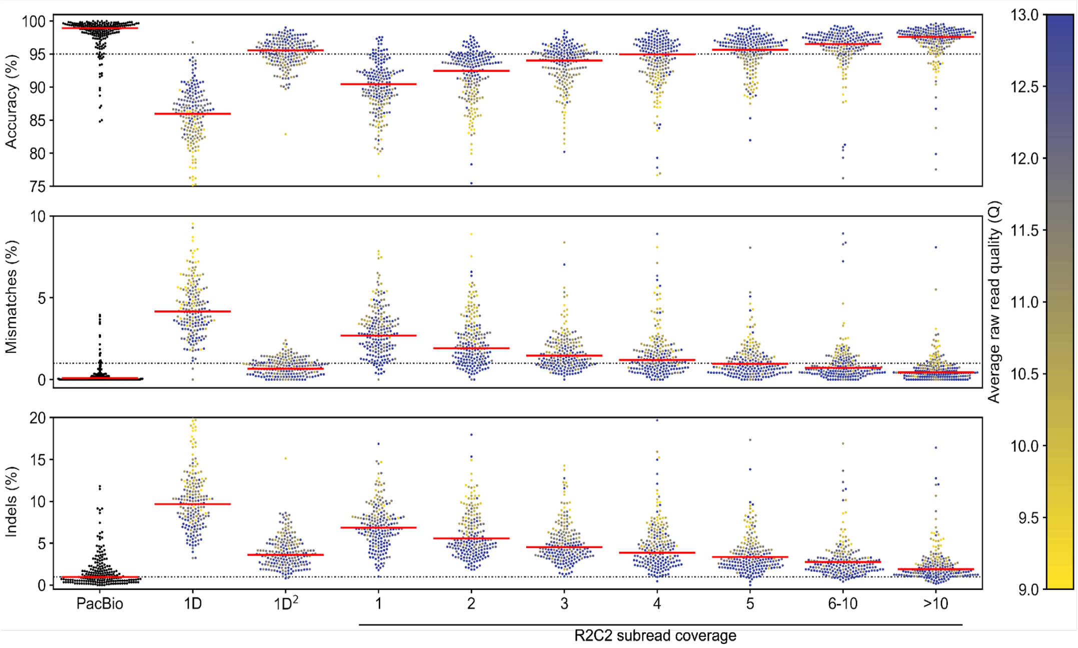


Fig. 0.5 Increase R2C2 subreads decreases indels and mismatches.
Figure adapted from (Volden et al., 2018). PacBio Isoseq, standard ONT 1D, and 1D^2 are compared with R2C2 at different subread coverages. Read accuracy is determined by minimap2 alignments to SIRV transcripts. Median accuracy is shown as a red line. Accuracy distribution is shown as a swarm plot of 250 randomly subsampled reads. Average raw read quality of ONT reads is indicated by the color of the individual points.

e) Data analysis

The goal of long-read transcriptome analysis is two-fold. First, it aims to identify all transcript isoforms present in a sample, then quantify their expression (ideally in an allele specific manner).

In contrast to short-read RNA-seq where bioinformaticians have spent the last decade creating a large number of tools for data analysis steps including read alignment, expression quantification, and transcriptome assembly, the tools for long-read analysis are still in their infancy. Although long-read technology circumvents many of the bioinformatic assembly challenges of short-read data, error-prone long-read data has created its own new set of challenges. These challenges have necessitated new algorithms for the efficient analysis of longer reads.

Nevertheless, interest among bioinformaticians towards long-read technology is steadily increasing and off-the-shelf tools to analyze long reads are being developed and published. Below is an overview of what the current state of tools is and what we perceive the outstanding challenges in the long-read cDNA field are.

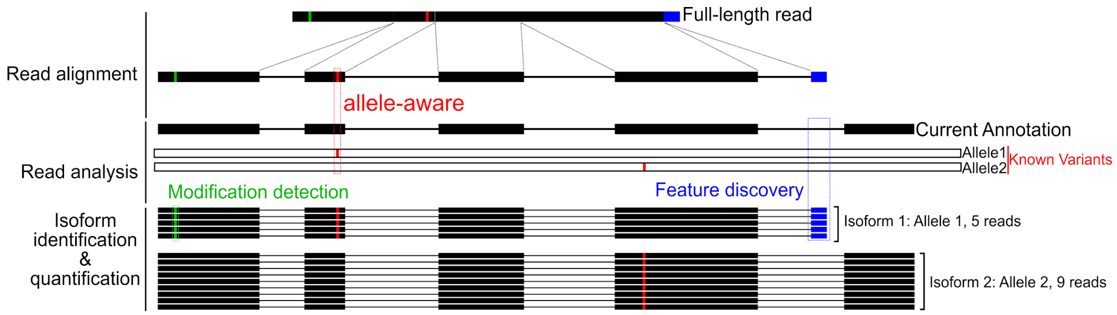


Figure 0.6: Analysis challenges of long-read full-length sequencing. A simplified schematic shows the steps required to extract information out of long-read sequencing data. Each read has to be aligned, ideally in an allele-aware manner to the genome it originated from. Read alignments then have to be analyzed to identify RNA modifications as well as new isoform features that are missing in the current transcriptome annotation. For each allele, reads then have to be grouped into isoforms which allows isoform identification and quantification. For real data sets, all these steps have to take into account the often-substantial rates of sequencing errors and incomplete reads in long-read sequencing. These will complicate all steps of the analysis.

i) Aligning long-read data

Aligning reads to a genome sequences is at the core of most transcriptome analysis (Fig. 0.6). Luckily there are several good options available for the spliced alignment of noisy long reads. The *GMAP* (T. D. Wu and Watanabe 2005) and *BLAT* (Kent 2002) aligners, originally developed for the alignment of ESTs perform surprisingly well for aligning noisy long reads. However, just like the PacBio developed *BLASR* (Chaisson and Tesler 2012) aligner, they are simply too slow for the effective analysis of millions of reads. The recently released *minimap2* (H. Li 2017) aligner seems to address the issue of speed while maintaining alignment accuracy and has quickly been adopted amongst the ONT community. The only trade-off we have observed (however not systematically investigated) is that *minimap2* – potentially due to the relatively large default seed size of 15nt – seems to lack sensitivity when aligning reads to very short terminal exons. We hope that future improvements in long-read accuracy will allow alignment algorithms to "dial in" that trade-off between avoiding spurious short alignments and detecting even the shortest of potentially un-annotated terminal exons.

ii) Isoform identification

Transcriptome annotation includes the identification of new gene features as well as how these new features are combined with known features into isoforms (Fig. 0.6). This is where long-read transcriptome sequencing holds the largest promise. However, the tools available for the identification isoforms from long read data are still in their infancy.

While PacBio supplies the *IsoSeq3* analysis pipeline for the analysis of their cDNA CCS reads, previous work indicates that this pipeline tends to over-report potential isoforms (Tardaguila et al. 2017). There currently exist three pipelines for the analysis of ONT direct cDNA or direct RNA sequencing data. Both *Pinfish* released by ONT and *FLAIR* released by the Brooks lab at UCSC are intended for regular 1D ONT data and deal with the high error-rate in different ways. Of these two pipelines, only *FLAIR* has been used in a publicly available manuscript [ref] and deals with inaccurate ONT reads by using short-read Illumina reads to correct splice junctions and identifies and quantifies isoform data; however, it does not use nanopore reads for de novo splice site detection and relies on annotation and short-read data (Workman et al. 2018).

Specifically designed for the analysis of R2C2 reads, the *Mandalorion* pipeline developed by our lab at UCSC takes advantage of the higher accuracy of R2C2 reads to identify and quantify isoforms without the need for Illumina data, while also identifying new gene features and isoforms (Volden et al. 2018).

One consideration when identifying isoforms is how to deal with raw data containing molecular biology artifacts. First and foremost, this includes the amplification of either fragmented RNA or genomic DNA. While, ideally, these artifacts should be minimized during sample preparation, any pipeline should be equipped to recognize potentially incorrect isoforms stemming from them. Tools like *Sqanti* which can detect these types of artifacts can serve as quality control for future isoform identification pipelines (Tardaguila et al. 2018).

iii) Isoform quantification

Quantifying and performing differential expression analysis of transcript levels on the isoform instead of the gene level brings with it a large set of new challenges.

First, it will be a challenge to decide at which point a known and a newly identified isoform should be treated as the same or equivalent isoforms. Containing different splice sites surely differentiates isoforms, but whether different TSSs that are only 3 nucleotides apart and reside within the 5' UTR differentiate isoforms is not at all clear.

FLAIR and *Mandalorion* deal with this by analyzing all samples that have to be compared at the same time to create a shared list of isoforms. This creates large computational overhead because adding a single sample to a data set requires the reanalysis of the entire data set.

Second, it will be a challenge to systematically differentiate allele-specific isoform expression (Fig. 0.6). To differentiate alleles, we will need accurate and phased information of sequence variants differentiating the haplotypes present in a sample, because extracting this information from error-prone long-read transcriptome data is inherently suboptimal. However, if sequence variants are known, tools like *HapCUT2* can be used to assign full-length cDNA to parental alleles (Edge, Bafna, and Bansal 2017). This in principle allows for allele-specific expression analysis.

We are, however, optimistic that approaches that sort aligned reads based on variants are only a temporary solution. In the future, it is likely that alignment algorithms will be able to take advantage of fully diploid genome sequences during alignment to immediately align reads to the haplotype they originate from. Then, ideally, future tools will identify allele-specific isoforms based on these alignments and quantify them using approaches similar to *RSEM* which uses expectation maximization to accurately quantify expression using short read data (B. Li and Dewey 2011).

iv) Modification detection

RNA transcripts are known to host a variety of base modifications than genomic DNA. Except for A-to-I modifications which are read by the RT enzyme as G and therefore appear in cDNA, RNA modification cannot be detected by standard cDNA sequencing as performed by Illumina, PacBio and ONT (Park et al. 2012).

Direct RNA sequencing, which is now possible on ONT sequencers, therefore holds great potential for modification discovery (Fig. 0.6). To realize this potential both computational and experimental workflows will need to be developed and improved. Although anecdotal evidence exists that modification information can be extracted from ONT base and raw data, no experimental and computational workflows exist yet to systematically establish and validate the detection of the large variety of modifications present in RNA (Workman et al. 2018). Furthermore, improvements to experimental workflows will have to reduce the RNA input requirements which currently limit direct RNA sequencing to large samples or cell lines.

The detection of DNA modifications using raw PacBio data may serve as a cautionary tale here (Flusberg et al. 2010). While the detection of methylated bases was shown to be possible using raw PacBio data, this approach never managed to compete with Illumina-based bisulfite sequencing for methylation detection. However, direct RNA sequencing has the potential to detect RNA modifications for which currently no other sequencing assay exist and might therefore fill a unique niche in the genomic toolset.

Conclusion

There is little doubt in my mind that full-length transcriptome sequencing using long-read technologies is the future of transcriptome annotation because it has too many inherent advantages over short-read approaches.

A single long read covering a full-length transcript can determine its transcription start site (TSS), all splice sites, and polyA site, thereby immediately identifying the isoform the transcript represents. In contrast, regular short-read RNA-seq protocols rarely detect TSSs and polyA sites and usually only cover a subset of splice-sites, leaving the researcher with a large computational problem when trying to identify isoforms which often has no clear solution.

We are confident that in the next few years, by addressing the challenges we describe here, long-read sequencing will make high-quality transcriptome annotations readily achievable within a reasonable budget. This will be of particular interest to researchers working on organisms that haven't attracted the attention of large consortia. Going forward, using 10X Genomics or Drop-seq approaches paired with long-read sequencing technology would allow for the amplification and sequencing of full-length cDNA from single-cell organisms to generate detailed isoform-level transcriptome annotations. Processing tens of thousands of cells this way could help generate an atlas of cells and would vastly expand our knowledge of the diversity of eukaryotic transcriptomes.

Chapter 1

Tn5Prime, a Tn5 based 5' Capture Method for Single Cell RNA-seq

Charles Cole[1,3*] Ashley Byrne[2,3*+], Anna E. Beaudin[1,4], E. Camilla Forsberg[1,5], Christopher Vollmers[1,6]

[This chapter is adapted from a publication **Tn5Prime, a Tn5 based 5' Capture Method for Single Cell RNA-seq** (Cole, Byrne et al., 2018, *Nucleic Acids Research*)]

* The first two authors should be regarded as joint First Authors.

1) Department of Biomolecular Engineering, University of California Santa Cruz, CABB
2) Department of Molecular, Cellular, Developmental Biology, University of California Santa Cruz, CA
3) The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors
4) Current Address: Department of Molecular and Cell Biology, School of Natural Sciences, University of California-Merced, Merced, CA, USA
5) Institute for the Biology of Stem Cells, University of California Santa Cruz, CA
6) Corresponding author. Email: vollmers@ucsc.edu

+ The author contribution was generating data, experimental design and writing the manuscript.

Abstract

RNA-seq is a powerful technique to investigate and quantify entire transcriptomes. Recent advances in the field have made it possible to explore the transcriptomes of single cells. However, most widely used RNA-seq protocols fail to provide crucial information regarding transcription start sites. Here we present a protocol, Tn5Prime, that takes advantage of the Tn5 transposase based Smartseq2 protocol to create RNA-seq libraries that capture the 5' end of transcripts. The Tn5Prime method dramatically streamlines the 5' capture process and is both cost effective and reliable. By applying Tn5Prime to bulk RNA and single cell samples we were able to define transcription start sites as well as quantify transcriptomes at high accuracy and reproducibility. Additionally, similar to 3' end based high-throughput methods like Drop-Seq and 10X Genomics Chromium, the 5' capture Tn5Prime method allows the introduction of cellular identifiers during reverse transcription, simplifying the analysis of large numbers of single cells. In contrast to 3' end based methods, Tn5Prime also enables the assembly of the variable 5' ends of antibody sequences present in single B-cell data. Therefore, Tn5Prime presents a robust tool for both basic and applied research into the adaptive immune system and beyond.

Introduction

As the cost of RNA-sequencing has decreased, it has become the gold standard in interrogating complete transcriptomes from bulk samples and single cells. RNA-seq is a powerful tool to determine gene expression profiles and identify transcript features like splice-sites. However, standard approaches lose sequencing coverage towards the very end of transcripts. This reduced coverage means that we cannot confidently define the 5’ ends of mRNA transcripts which contain crucial information on transcription initiation start sites (TSSs) and 5’ untranslated regions (5’UTRs). Analyzing TSSs can help infer the active promoter landscape, which may vary from tissue to tissue and cell to cell. Analyzing 5’UTRs, which may contain regulatory elements and structural variations can help infer mRNA stability, localization, and translational efficiency. Identifying such features can help elucidate our understanding of the molecular mechanisms that regulate gene expression.

The loss of sequencing coverage towards the 5’ end of transcripts is often attributed to how sequencing libraries are constructed. For example, the widely used Smart-seq2 RNA-seq protocol, a powerful tool in deciphering the complexity of single cell heterogeneity (Picelli, Faridani, et al. 2014; Treutlein, Brownfield, et al. 2014; Darmanis et al. 2015), features reduced sequencing coverage towards transcript ends. This lost information is a result of cDNA fragmentation using Tn5 transposase. Several technologies have tried to compensate for the lack of coverage by specifically targeting the 5’ ends of transcripts. The most notable methods include cap analysis of gene expression (CAGE), NanoCAGE, and single-cell tagged reverse transcription sequencing (STRT) (S. Islam et al. 2011, 2014; Salimullah et al. 2011; Shiraki et al. 2003). CAGE uses a 5' trapping technique to enrich for the 5'-capped regions by reverse transcription (Shiraki et al. 2003). This technique is extremely labor intensive and involves large amounts of input RNA. The NanoCAGE and STRT methods target transcripts using random or polyA priming and a template-switch oligo technique to generate cDNA (S. Islam et al. 2011; Salimullah et al. 2011). While NanoCAGE can analyze samples as low as a few nanograms of RNA, and STRT can be used to analyze single cells, they both require long and labor-intensive workflows including fragmentation, ligation, or enrichment steps. These workflows can become costly and labor intensive, making it difficult to interrogate complex mixtures of cells like those found in the adaptive immune system or cancer.

New droplet based high-throughput single cell RNAseq approaches like Drop-Seq and 10X Genomics Chromium platform can process thousands of cells but require intricate or expensive proprietary instrumentation. Importantly they are primarily focused on the 3' end of transcripts due to integrating a sequencing priming site onto the oligodT primer used for reverse transcription. By losing information of the 5' end almost entirely, these approaches are not capable of comprehensively analyzing cells of the adaptive immune cells which express antibody or T cell receptor transcripts featuring unique V(D)J rearrangement sequence information on their 5' end. While 10X Genomics has recently introduced their new Single Cell V(D)J solution platform to address this we have yet to discover how well this new method works.

To overcome this lack of easy-to-implement, inexpensive, and high-throughput single cell 5’ capture methods, we chose to modify the Smart-seq2 library preparation protocol, which is relatively cost-effective and simple with features of STRT which captures 5’ ends effectively. Here we describe a robust and easily implemented method called Tn5Prime that performs genome-wide profiling across the 5’ end of mRNA transcripts in both bulk and single cell samples. The protocol is based on integrating one sequencing priming site into the template switch oligo used for reverse transcription and subsequently tagmenting the resulting amplified cDNA by Tn5 enzyme loaded with an adapter carrying the other sequencing priming site. This combination allows for the construction of directional RNAseq libraries with one read anchored to the 5’ end of transcripts without the need for separate fragmentation, ligation, and, most importantly, enrichment steps. Additionally, by incorporating cellular identifiers into the template switch oligo makes it conducive for pooling samples after reverse transcription, thereby increasing throughput and reducing cost. Finally, data produced by this novel approach allows for the identification of transcription start sites, the quantification of transcripts, and the assembly of antibody heavy and light chain sequences from single B cells at low sequencing depth.

Results

Construction of Tn5Prime libraries

Tn5Prime libraries can be constructed from either purified total RNA or single cells sorted by Fluorescence-activated cell sorting (FACS) into multiwell PCR plates. Tn5Prime libraries create a directional paired-end Illumina RNAseq library with read 1 anchored to the 5’ end of transcripts. Directionality and read 1 anchoring is achieved through the use of our modified template-switch oligo and custom Tn5 enzyme. After the addition of reverse transcriptase to total RNA or cell lysate, first-strand synthesis occurs using a modified oligo-dT and a template-switch oligo (TSO) containing a partial Nextera A adapter sequence and, optionally, a cellular index sequence (Supplementary Table S1.1, Fig. 1.1A). During reverse transcription, the oligo-dT serves as a primer at the 3’ polyA tail of mRNA transcripts, while the sequence of the partial Nextera A template-switch oligo is attached to the 3’ end of the synthesized cDNA corresponding to the 5’ end of transcript sequences. After reverse transcription, samples with non-overlapping cellular indexes can be pooled. The cDNA product is then amplified using a complete Nextera A primer and a primer complementary to the modified 5’ end of the oligo-dT. After amplification, the cDNA product will contain a complete Nextera A adapter including Illumina indexes. At this point, samples that contain the non-overlapping Illumina indexes can be pooled. By pooling after reverse transcription and PCR amplification, we can dramatically reduce the workflow complexity and reagent usage.

Next, Tn5 transposase, loaded only with a partial Nextera B adapters, fragments the cDNA and attaches the partial Nextera B adapters to the cDNA in a single reaction. The cDNA fragments are then amplified using a universal A primer and a Nextera B primer that primes off the partial Nextera B adapter sequences attached by the Tn5 enzyme. The final product is compatible with the Illumina platform by containing the complete Nextera A and Nextera B adapters. Libraries are then ready to be size selected and quantified prior to sequencing. At this point, no enrichment step is necessary, as only molecules containing both Nextera A and B adapters will be targeted for sequencing. Since only the TSOs associated with the 5' end of transcripts contains Nextera A adapters, read 1 of all read pairs in the sequencing reaction begins at these 5' ends and extends into the transcript body, thereby identifying transcription start site and directionality (Fig. 1.1A-C). Read 2 is distributed throughout the gene body, as each location represents the random insertion of Nextera B adapters by Tn5 and library size selection (Fig. 1.1B, C)

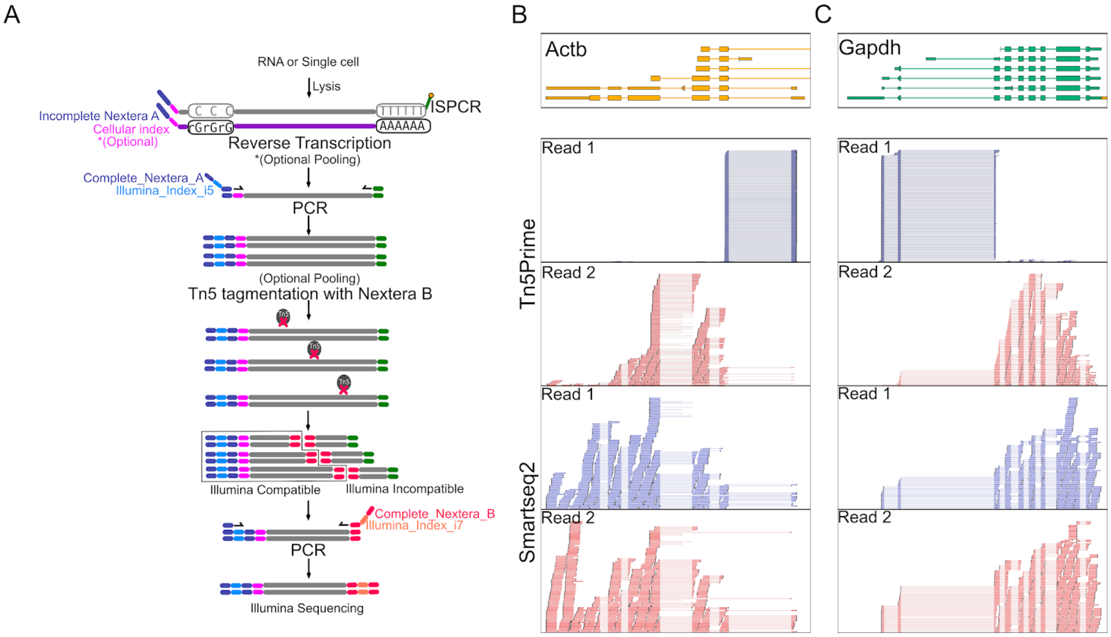


Fig. 1.1 Tn5Prime Library construction and 5’ capture
A.) Schematic of the Tn5Prime library construction. No enrichment steps are required to generate a library that captures the 5’ end of transcripts. B.) Read alignment plots comparing 5’ end capture by Tn5Prime to random fragmentation by Smartseq2 using lymphoblast cell line GM12878. A total input of 50 ng of RNA was used. Individual alignments for the first (Read1, blue) and second (Read2, red) read of each read pair are shown. Read1 density is shown for both library types as a histogram (blue). Gene models are shown on the top panel (Color indicates transcriptional direction.)

Creating and analyzing Tn5Prime data of GM12878 cell line RNA

To evaluate whether our Tn5Prime protocol consistently identifies the 5’ end of the transcript we first performed low coverage RNAseq of total RNA of GM12878 cultured lymphoblast cells. We performed a side-by-side comparison of our protocol with a modified version of the Smart-seq2 (Picelli, Faridani, et al. 2014; Byrne et al. 2017) (see Methods) protocol using the same starting material. Using the HiSeq2500 platform (Illumina) we obtained 570805 and 453761 raw read pairs for two replicate Tn5Prime libraries. We next obtained 1094530 raw read pairs from the Smart-seq2 library. Adapter sequences and low quality reads were removed using Trimmomatic (Bolger, Lohse, and Usadel 2014). In the Tn5Prime replicates, 92.51% and 92.62% of the trimmed and filtered reads mapped uniquely to the human genome using the STAR alignment tool (Dobin et al. 2013), surpassing the Smart-seq2 protocol at 88.50%. The uniquely aligned reads from the TN5Prime replicates collectively had a redundancy of 1.34. This high unique alignment percentage indicates that our Tn5Prime protocol produces libraries of high complexity.

Detecting Transcription Start Sites using Tn5Prime

We analyzed the read distribution across transcripts both visually and systematically to determine the 5' specificity of our protocol. Visual inspection found that while Smart-seq2 reads are distributed across the entire body of genes, Tn5Prime reads follow two distinct patterns: First, the start of the read 1 is anchored to the transcription start site. Second, the start of read 2 is variable and likely dependent on size selection during library preparation (Fig. 1.1B-C). Next, systematic analysis was based on mapping the start of read 1 to identify putative transcription initiation start sites (TSSs). To test our ability to identify TSSs, we compared our Tn5Prime data to the Gencode genome annotation and CAGE data which was generated from the same GM12878 cell line from the ENCODE project. We identified putative TSSs by calling peaks enriched from the start of read 1 in our Tn5Prime data (see Methods). We found that 89.7% of the 17,853 peaks fell within TSSs (0-25bp upstream) with the vast majority of them falling near promoter regions (26bp-1000bp upstream) or

5’UTRs (Fig.1.2A). Next, we subsampled the CAGE data to levels similar to the Tn5Prime data and called peaks in the same manner. We found 73% of the 17,853 Tn5Prime peaks fell within 25bp to the nearest of 27,526 CAGE peaks, indicating high concordance between the two approaches (Fig. 1.2B). Tn5Prime peaks (3,746) that were not within 25bp of a CAGE peak contained far less sequencing reads on average than those within 25bp of a CAGE peak. These results indicate that these transcripts might be expressed at lower levels and show more variance between the Tn5Prime and CAGE datasets (Fig. 1.2B). Next, as a comparison we looked at our GM12878 data generated using the Smart-seq2 method in the same way. We found that 7.9% of the 23,451 peaks called based on the Smart-seq2 fell within TSSs (0-25bp upstream) (Fig. 1.2C). Further, we found 10.4% of the 23,451 peaks fell within 25bp to the nearest CAGE peaks (Fig. 1.2D). This showed that in contrast to the Smart-seq2 method our TN5Prime approach effectively identified putative TSS sites. Ultimately, this data suggests that our Tn5Prime protocol is equivalent to the gold standard CAGE technique in targeting transcription start sites.

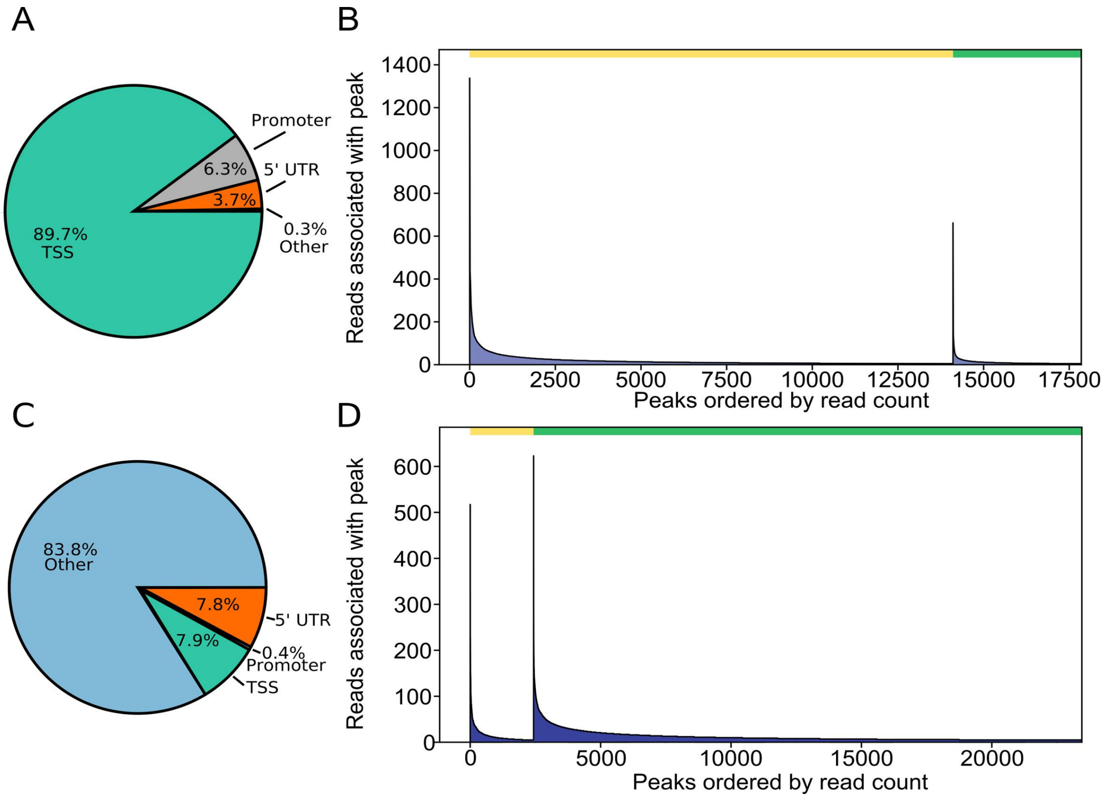


Fig. 1.2 Tn5Prime peaks are highly concordant with GENCODE annotation and CAGE peaks

Peaks were detected from sequencing reads produced by Tn5Prime and Smart-seq2 libraries generated from total GM12878 RNA. (**A** and **C**) Tn5Prime (A) and Smart-seq2 (C) were matched to features in the Gencode annotation and the feature they matched are shown as a pie chart. (**B** and **D**) Tn5Prime (B) and Smart-seq2 (D) peaks were matched to CAGE peaks. The yellow bar on top indicates the peaks within 25 bp and the green bar indicates all other peaks. Peaks in each were rank sorted according to their read coverage and shown as a histogram.

Quantifying the Transcriptome using Tn5Prime

After validating the ability of Tn5Prime to detect transcription start sites, we next wanted to examine whether it is capable of transcript quantification. To determine whether our Tn5Prime method is quantitative we compared GM12878 data generated from four different protocols: Tn5Prime, Smart-seq2 data generated by our lab, as well as CAGE and RNA-seq data produced by the Encyclopedia of DNA Elements (ENCODE) project (Fig. 1.3). We used the Tn5Prime data mentioned in the previous section and generated the Smart-seq2 data on the same cell line as described by (Picelli, Faridani, et al. 2014). We performed replicates using the Tn5Prime protocols to define overall reproducibility and accuracy. Based upon our results, transcript quantification by Tn5Prime replicates showed extremely high correlation with a Pearson correlation coefficient of r=0.97 (95% C.I. 0.97-0.97). Quantification by Tn5Prime also correlated very well with Smart-seq2 with a Pearson r of 0.87 (95% C.I. 0.86-0.87). Tn5Prime and Smart-seq2 data were comparable with ENCODE RNA-seq and CAGE data (Fig. 1.3), indicating that the Tn5Prime protocol is equivalent to the conventional Smart-seq2 method in measuring transcript abundance. Together, these data show that Tn5Prime can accurately identify transcription start sites and quantitatively measure transcript abundance.

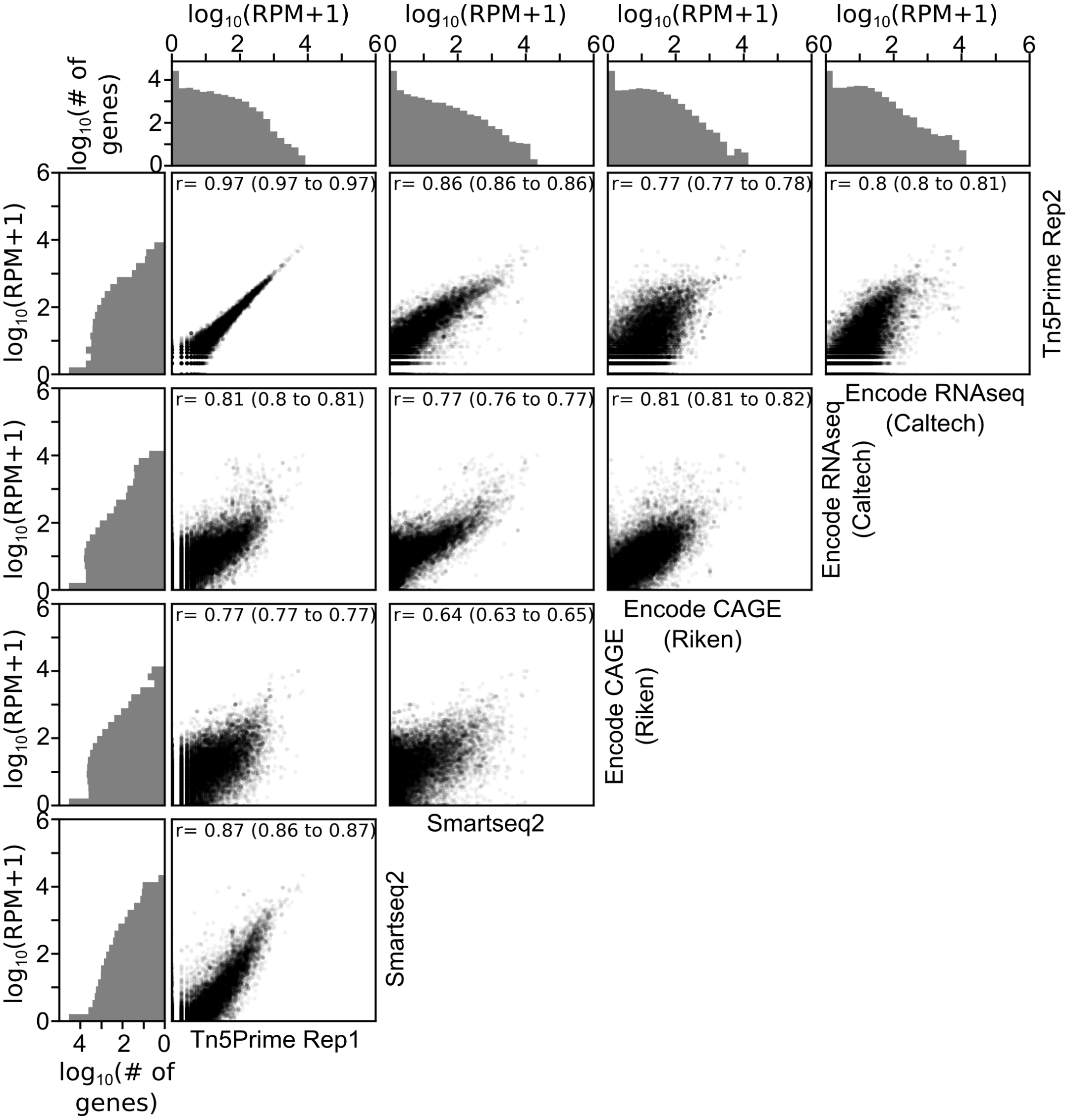


Fig 1.3. Tn5Prime quantifies transcriptomes accurately and reproducibly. Pairwise correlations of transcript levels between Tn5Prime, Smartseq2, ENCODE CAGE and ENCODE RNAseq experiments using GM12878 cell line are shown as scatter plots. A total of 50 ng of input RNA was used. Each transcript is shown as a black dot with an opacity of 5%. Distribution of transcript levels is shown on the outside of the plots in grey histograms.

Transcript quantification and start site localization in single B cells.

As the Tn5Prime protocol is based on the same cDNA amplification strategy as the Smart-seq2 protocol, we expected it capable of generating sequencing libraries from single cells. Indeed, we successfully generated single cell libraries using the Tn5Prime protocol from primary murine B-lymphocytes (B2 cells; IgM+B220+CD5-CD11b-) (n=12) isolated from the peritoneal cavity. We generated between 17,534-93,429 2x300 bp read pairs per cell using the Illumina MiSeq of which 62% passed quality filtering. Of the filtered reads, an average of 91.48% uniquely mapped to the mouse genome. The high alignment percentage indicates we are able to generate high quality libraries from single cells using our Tn5Prime. Despite the very low total number of read pairs we collected, we still detected 339 expressed genes per cell on average. Although these numbers may seem low, they are in line with previous published single B cell RNAseq studies (Gierahn et al. 2017; Zheng et al. 2017; Jaitin et al. 2014). Also, it is known that B cells can show transcriptional heterogeneity depending upon their cell state (Y. L. Wu et al. 2017). Among the genes expressed in many of the single cells were genes corresponding to B cell function, including CD19, CD79a and components of the MHC complexes (Supplementary Fig. S1.1). This data indicates that we were able to effectively identify cell type specific genes.

Analysis of 192 Single CD27^{high} CD38^{high} Human B Cells

After successfully testing our Tn5Prime method on single mouse B cells, we next wanted to develop a multiplex approach capable of evaluating hundreds of human single cells. To this end, we FACS sorted 192 single B cells into individual wells of 96 well plates using the canonical surface molecules CD19, CD27 and CD38 to sub-select the plasmablast subpopulation (Supplementary Fig. S1.2). Plasmablasts are one of the most widely studied B cell populations and are frequently monitored after vaccination or infections by flow cytometry. The plasmablast cell compartment can be defined primarily by looking at high levels of surface markers CD27 and CD38, however memory B cells may also express these markers, albeit at lower levels, making it difficult to parse out these two populations. Therefore, analyzing these cell types at the single cell level should help further delineate these populations.

Our multiplex strategy entails inserting cellular indexes into the template switch oligo allowing the libraries to be pooled after reverse transcription. This streamlines our method and increases our throughput by decreasing the PCR and Tn5 reactions required. Using our multiplexing strategy, we generated Tn5 libraries for 192 single B cells using 192 RT reactions, 24 PCR reactions and 24 Tn5 reactions. Although this was not performed, library pools carrying distinct Illumina sample indexes could have been further pooled following PCR to reduce the numbers of Tn5 reactions from 24 to 2. The entire Tn5Prime library preparation workflow for hundreds of cells can be completed in two days.

We generated 194,553,648 150 bp paired end reads total. To determine gene expression for each cell, reads were assigned to one of 192 single cells based on its Illumina index reads and by comparing the sequence of the first 8 bases of read 1 to the cellular index sequences. 91% of the 194,553,648 150bp paired end reads were successfully assigned to one of the 192 single B cells. 90.75% of cell-assigned reads were successfully aligned to the human genome using the STAR alignment tool with a median of 74.59% or 505,665 of cell-assigned reads being uniquely assigned to an annotated gene. Each cell expressed a median of 534 genes. We then compared the number of genes detected by Tn5Prime and modified Smart-seq2. To this end, we sequenced 13 Smart-seq2 B cells libraries to a median depth of 275,762 reads uniquely aligned to genes. When subsampled to the median Smart-seq2 read depth of 275,000 reads Tn5Prime detected a median 409 genes while Smart-seq2 detected 910. While detecting less genes than Smart-seq2, the Tn5Prime method is comparable to other high-throughput single cell methods like MARS-seq (Jaitin et al. 2014) (Median of 671 genes per B cell), 10X Genomics (Zheng et al. 2017) (Median of 478 genes per B cell), and Seq-well (Gierahn et al. 2017) (Median of 874 genes per B cell).

Overall, of the 58234 annotated genes in GENCODE, 5414 genes had at least one read per cell on average among the 192 B cells analyzed with Tn5Prime. The median redundancy for each cell is 13.92 which means that, on average, each uniquely aligned cDNA fragment was sequenced 13.92 times. This indicates that the libraries were sequenced exhaustively.

Detecting Transcription Start Sites in single CD27^{high} CD38^{high} B cells using Tn5Prime

To determine if transcription start site specificity is maintained within the single cell data, read 1 start distribution was compared to annotated transcription start sites found in the ENCODE and CAGE datasets. By calling peaks, we found that our single cell results were able to maintain transcription start site specificity, with peaks predominantly falling within the annotated transcription start sites with 92.4% of the peaks falling within TSSs (Fig. 1.4A-B). In addition to the transcription start site, the directionality of transcription can be inferred due to our custom template switch oligo incorporating a forward-read priming site to the 5' region of the transcript which is an advantage over many other single cell RNAseq protocols (Fig. 1.4C-D).

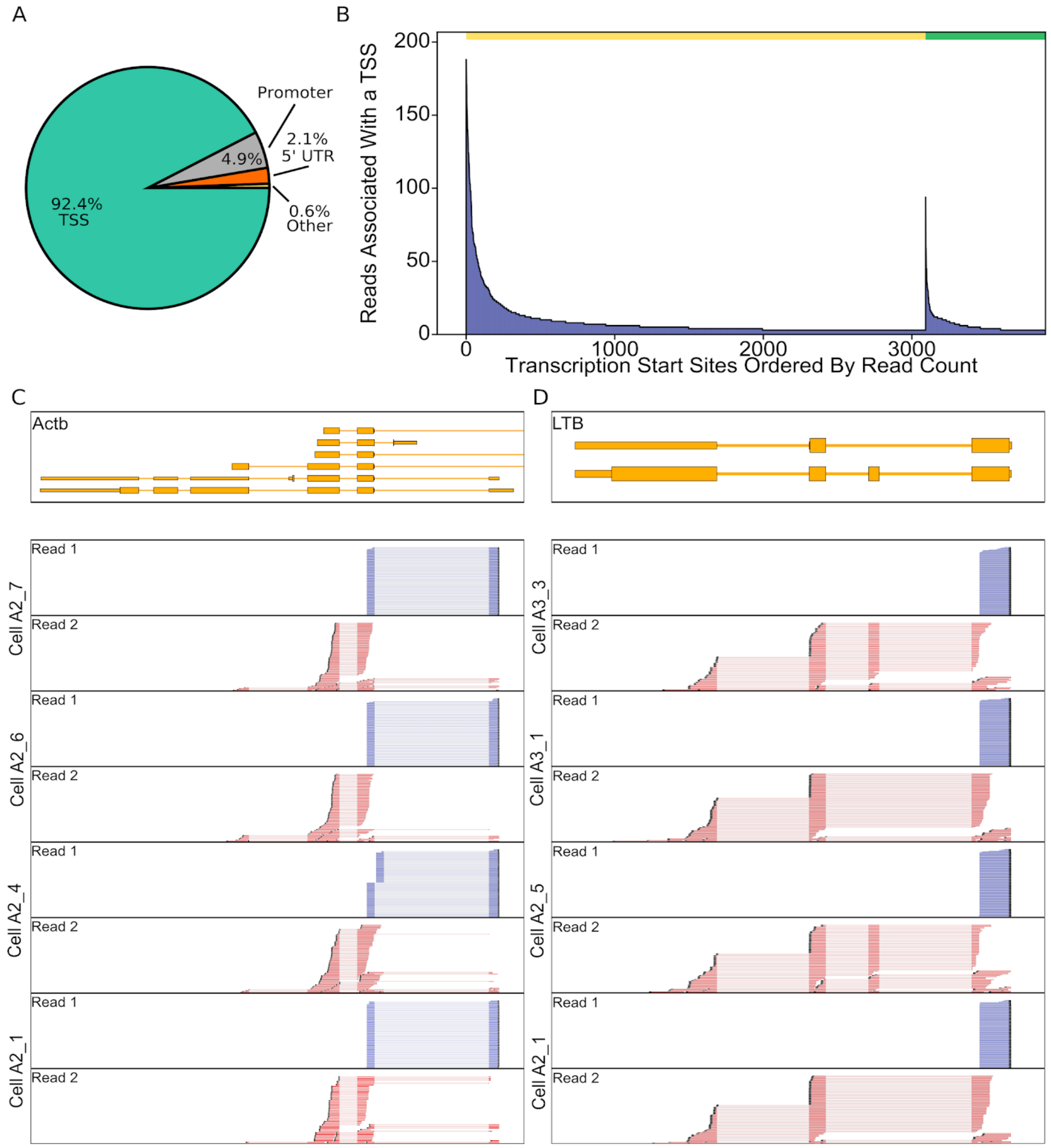


Fig 1.4. Transcription start sites detected in single CD27^{high} CD38^{high} B cells

A) CD27^{high} CD38^{high} Tn5Prime peaks matching features in the Gencode annotation. TSS = on or < 25bp behind the start of an annotated GENCODE gene, 5'UTR = inside 5' UTR, Promoter = between 26 -1000bp behind start of annotated gene. B) Tn5 peaks shown in two groups. Group 1 contains all peaks within 25bp of a peak identified in the complete RIKEN CAGE peak Human peak database and group 2 contains all other peaks. Peaks were sorted by the number of cells associated with that peak in CD27^{high} CD28^{high} B cells and displayed in figure 5a. The yellow bar indicates peaks within 25bp, and the green bar indicates all other peaks. C, D) Genome Browser view of reads of Actb (C) and LTB (D) genes. In addition to TSS information, read alignments also show differential isoform usage between cells.

Detecting Subpopulations within CD27^{high} CD38^{high} B cells using Tn5Prime

Since separating memory B cells and plasmablasts by FACS based on surface markers can be challenging, especially when the adaptive immune system is unperturbed, we wanted to see whether we could do so post-sorting using their gene expression profiles. Cells outside more than three median absolute deviations from the median for percent alignment, Mitochondrial transcript percentage, or number of detected genes were marked as outliers and eliminated prior to normalization of transcript counts (Supplementary Fig. S1.3). After normalizing raw gene expression counts and removing non-recombined and therefore non-applicable antibody gene segments from the annotation (Lun, Bach, and Marioni 2016), we clustered the remaining 159 sorted B cells using t-SNE dimensional reduction. The clusters were robust when the data was subsampled to 100,000 reads per cell (Supplementary Fig. S1.4). We then identified genes that showed significant differences between the two clusters. We detected 411 genes with significant changes including J-chain, LTB (Lymphotoxin Beta), XBP-1 (X-box binding protein 1), HSPA5 (Heat-shock protein family A), and MZB1 (Marginal Zone B1). We also found genes HLA-DRA, HLA-DRB5, and HLA-DPB1, which encode for the alpha and beta chains of the MHC II (Major Histocompatibility Complex II) to be differentially expressed (Supplementary Table S1.2). The J-chain was upregulated in cluster 2 and is involved in antibody secretion of IgM and IgA (Lamson and Koshland 1984) (Fig. 1.5). XBP-1, MZB1 and HSPA5 were upregulated within cluster 2 and are known targets of BLIMP-1.

BLIMP-1 and XBP-1 are known to be essential in plasmablast differentiation (Supplementary Fig. S1.5) (Minnich et al. 2016; Nutt et al. 2015). LTB was downregulated in cluster 2 and has been shown to be downregulated upon B cell activation (Zhu et al. 2004) (Fig. 1.5). HLA-DRA, HLA-DRB5, and HLA-DPB1 were downregulated in cluster 2, indicating less MHC II presentation to T cells, which is indicative of plasma cells and plasmablasts (Calame, Lin, and Tunyaplin 2003). Together, this suggests that cluster 2 does represent activated plasmablasts, which are known to secrete more antibodies and display less MHC II than the memory B cells represented in cluster 1.

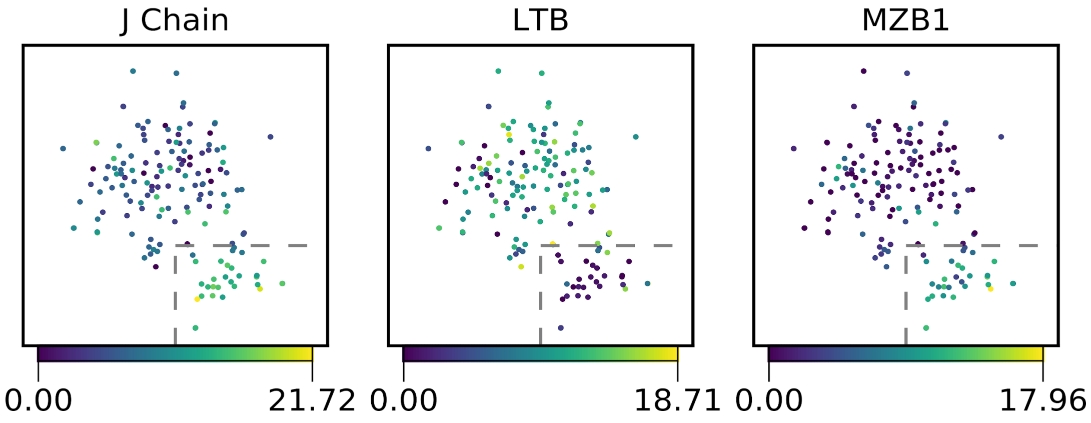


Figure 1.5. Clustering of CD27^{high} CD38^{high} B cells
159 B cells were divided into two populations by t-SNE dimensionality reduction (Maaten and Hinton 2008). In the three subplots, cells are colored based on their expression of example genes that were significantly differentially expressed between the two populations as determined by a multiple hypothesis testing corrected Mann-Whitney U tests. The cells inside the boxed area belong to cluster 2 and all other cells belong to cluster 1.

Assembly of antibody heavy and light chain sequences from single B cell Tn5Prime data

Ideally, we would not only want to identify plasmablasts based on their gene expression profile, but also determine their antibody sequences. Sequencing antibodies has been a long-standing challenge in B cell biology and antibody engineering because it requires the identification of unique pairs of rearranged antibody heavy and light chains for each cell. Current techniques rely either on the targeted amplification and sequencing of antibody heavy and light chain genes (Wrammert et al. 2008) in single cells or on the assembly of their sequences from non-targeted RNA-seq data (Canzar et al. 2017). As a result, our 5' capturing approach we could potentially provide antibody sequence information in addition to genome wide expression profiling, because the 5' region contains the unique V(D)J rearrangement of heavy and light chain transcripts.

To determine if our Tn5Prime protocol could be used for assembling antibody heavy and light chain sequences, we assembled whole transcriptomes using SPAdes (Bankevich et al. 2012). IgBLAST (Ye et al. 2013) was then used to identify transcripts containing V, D, and J gene segments rearranged in a productive manner. These transcripts were aligned on to Constant gene segments to identify isotype. The list of putative antibodies was then filtered for obvious cross-contamination and mis-assemblies (see Methods). In this way, we effectively determined heavy and light chain sequences and identify their unique pairings within single B cells (Fig. 1.6A).

Of the 192 B-cells we analyzed, we were able to assemble one heavy chain and one light chain to 117 B-cells. Of these 117 B-cells 46 cells had a Lambda light chain and 71 cells had a Kappa light chain. Five additional cells had one heavy chain and two light chains, 35 cells had no heavy chains but at least one light chain, and 35 cells had no heavy chains and no light chains. To determine the sequencing depth requirement for successful heavy and light chain assembly, subsampling was performed on the reads and the assembly and pairing analysis redone (Supplementary Fig. S1.6). We found 100,000 reads per cell was sufficient to assemble one heavy and one light chains for 91 of 117 B cells with successfully assembled chain pairs without subsampling.

We found that 101 of the 117 cells with paired heavy and light chains also passed all other quality filters and were clustered by t-SNE into the putative plasmablast and memory B cell clusters. This combination of single cell identity and paired antibody sequences allowed us to perform detailed analysis of differences in antibody usage and characteristics between those two populations. Firstly, the putative plasmablast population featured less IgM antibodies than the memory B cell population (19% IgM in plasmablasts vs 53% in memory B cells). Secondly, using IgBlast (Ye et al. 2013), we found that both heavy (Fig. 1.6B) and light chain sequences showed significantly higher levels of somatic hypermutation in plasmablasts than memory B cells (Heavy chain: median 8.0% vs 3.8% somatic hypermutation, two-sided Monte Carlo permutation test p-value=0.0081; Light chain: median 4.9% vs 2.7% somatic hypermutation, two-sided Monte Carlo permutation test p-value=0.0117). Thirdly, by counting and normalizing sequencing reads originating from antibody transcripts, we could determine and compare heavy and light chain expression in these two populations. Generally, light chains were expressed about 3-fold higher than heavy chains (Fig. 1.6C) with no significant difference between plasmablasts and memory B cells (two-sided Monte Carlo permutation test p-value=0.533). However, the percentage of all aligned sequencing reads that originated from antibody transcripts showed dramatic differences between plasmablasts and memory B cells. The median percentage of reads that originated from antibody transcripts was 23.5% in plasmablasts and only 2.2% in memory B cells (Fig. 1.6D) (Monte Carlo Permutation test two-sided p-value=0). In one plasmablast over 60% of all aligned sequencing reads originated from antibody transcripts indicating just how much of the plasmablast transcriptome can be dedicated to the production and secretion of antibodies. In summary, our analysis of antibody usage and characteristics showed that plasmablasts express more mutated and class-switched antibodies at much higher levels than memory B cells.

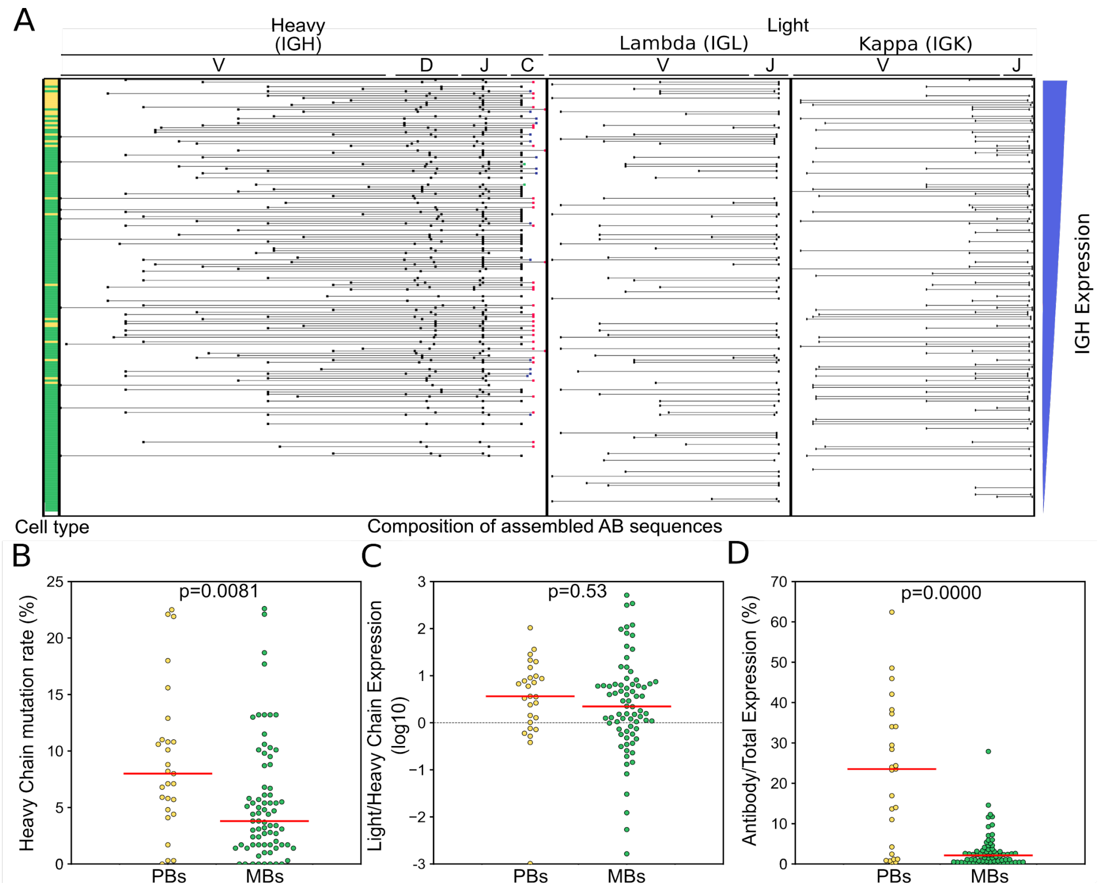


Figure 1.6. Assembling Antibody transcripts from Tn5Prime data

Antibody transcripts were assembled by generating complete assembled transcriptomes for each cell with SPADES and then using IGBLAST to search for transcripts with antibody features. Antibody transcripts for each cell were filtered for mis-assemblies and mis-annotations. Cells were sorted by the abundance of heavy chain transcripts in their Tn5Prime data and V (D) and J segment information for their heavy and light chains are shown in the schematic in the center. The putative cell type determined by clustering with t-SNE is indicated on the left. Yellow: plasmablasts, Green: Memory B cells. (B-D) Antibody usage and characteristics were compared between plasmablasts and memory B cells. Somatic Hypermutation rates (B), light to heavy chain expression ratios (C) and the percentage of all aligned sequencing reads that originated from antibody transcripts (D) were compared using dot plots. Yellow: plasmablasts, Green: Memory B cells. Medians are shown as red lines. All p-values are calculated using two-sided Monte Carlo permutation test with 10000 permutations.

Discussion

Here we present a novel method for the genome-wide identification of transcription start sites in bulk samples and single cells. The method combines aspects of both Smart-seq2 and STRT. By modifying template-switch oligos used during reverse transcription to carry one sequencing adapter and loading the other sequencing adapter on the Tn5 enzyme used for cDNA fragmentation we anchor the sequence priming sites for read 1 of an Illumina read pair to the 5' end of transcripts without the need for fragmentation, ligation, and enrichment steps. The resulting workflow is easy to implement and capable of generating hundreds of libraries within a day. An important feature of our Tn5Prime method is the option to integrate cellular indexes during reverse transcription and Illumina sample indexes during PCR before Tn5 tagmentation. This allows the pooling of samples early in the workflow and thereby reduces experiment complexity and reagent costs.

We validated the Tn5Prime protocol on both bulk RNA and single cells. First, using 5ng of total RNA from the GM12878 cell line, we yielded similar results as the ENCODE CAGE data with respect to the identification of transcripts start sites. However, the CAGE protocol used by the ENCODE consortium used several orders of magnitude more RNA. As the Smart-seq2 protocol is already widely used, we expect that the Tn5Prime assay with its similar workflow and low RNA input has the potential to become a valuable tool for transcriptome annotation and quantification in the RNA-seq toolbox.

In addition to the analysis of bulk samples, we show that our Tn5Prime method can be utilized for interrogating single cells, both human and mouse. The TSO-based multiplexing approach we implemented makes it possible to efficiently analyze thousands of cells, thereby increasing the throughput of plate based RNAseq library protocols in a manner that is straightforward and economical. While the Tn5Prime approach detects less genes than the Smart-seq2 approach, as determined by gene detection, this could be improved by increasing the amount of cDNA pooled for amplification (currently only ~60% of cDNA is used) as well as by using LNA bases in the Tn5Prime TSOs (Picelli, Faridani, et al. 2014), although the latter approach might affect 5' specificity (Harbers et al. 2013).

Our Tn5Prime approach interrogates the 5' ends of transcripts, thereby capturing the unique sequence information of adaptive immune system receptors expressed on B and T cells. These receptors are often hard to assemble due to their unique genomic rearrangement. Our data shows that by limiting sequencing reads to the 5' end of transcripts we can analyze both transcriptomes as well as paired antibody heavy and light sequences with the low sequencing coverage of ~100,000 reads per cell, thereby enabling the analysis of thousands of B cells in a single sequencing run. This approach should, without any modification, also be applicable to T cells to map rearrangement of the T cell receptors. This can provide novel insights into the composition of B and T cell malignancies as well as the state and composition of the adaptive immune system with regards to solid tumors. This sets Tn5Prime apart from general purpose high-throughput single cell library preparation methods like Drop-Seq, Seq-well, and 10X Genomics which target the 3' end of the transcripts making them incapable of interrogating antibody sequences. We are looking forward to more published data on the recently released 10X Genomics Single Cell V(D)J platform which should be able to, like Tn5Prime, investigate V(D)J expression and gene expression in parallel. Overall, determining per cell library preparation cost, required sequencing depth, and cell capture rate will help establish ideal use-cases for either Tn5Prime or 10X methods.

To highlight the power of our Tn5Prime approach we isolated 192 single human B cells from PBMCs using canonical plasmablast markers. Not only were we able to assemble paired antibody transcripts, but we succeeded in clustering the cells into two populations based on their gene expression profiles. The genes differentially expressed between those clustered suggested their putative cell types. Cells in the putative plasmablast cluster expressed more XBP-1, J-chain, HSPA5, and MZB1, which are all involved in either B cell activation or antibody production and secretion. Consistent with less antigen presentation, cells in the putative plasmablast cluster also expressed less MHC II transcripts including HLA-DRA, HLA-DRB5, and HLA-DPB1. Finally, MS4A1 (CD20) is also expressed less in the cells of the putative plasmablast cluster and is known to be downregulated in activated B cells. Overall, this clearly established that we could distinguish activated, antibody secreting plasmablasts from resting, antigen presenting memory B-cells; cell-types which are difficult to distinguish using conventional FACS analysis.

In addition to cell-types, we showed that Tn5Prime can be used to determine individual B cells' paired antibody sequences. Together, these data allowed us to compare antibody usage in plasmablasts and memory B cells, showing that plasmablast expressed higher levels of more mutated and class-switched antibodies. In addition to providing functional insight into cell populations, this information will make it possible to make informed decisions as to which antibody sequences could be further cloned and tested functionally for clinical, diagnostic, and research applications.

In summary, Tn5Prime is an RNAseq library construction protocol with a streamlined workflow that surpasses the economy and throughput of other plate-based protocols. While not reaching the throughput of droplet- and microwell-based protocols, it generates high quality data that enables the identification of transcription start sites and could be useful for analyzing 5' UTR features or help improve incomplete genome annotations. Finally, Tn5Prime is currently the highest throughput library preparation method that doesn't require proprietary instrumentation to comprehensively analyze the individual cells of the adaptive immune system by determining both paired adaptive immune receptor sequences and gene expression profiles.

Materials and Methods

Cell purification, RNA isolation and sorting

GM12878: RNA from 500,000 GM12878 cells was extracted using the RNeasy kit (Qiagen) according to manufacturer's instructions.

Murine B2 cells: Mice were maintained in the UCSC (University of California, Santa Cruz) vivarium according to IACUC (Institutional Animal Care and Use Committee) approved protocols. Single murine Ter119-CD3-CD4-CD8-B220$^+$ IgM$^+$CD11b$^-$ CD5$^-$ B2 cells were isolated from wild-type C57Bl/6 mice by peritoneal lavage and incubated with fluorescently-labeled antibodies prior to sorting. The following antibodies were used to stain B-cells: Ter119, CD3 (Biolegend; 145-2C11), CD4 (Biolegend; GK1.5), CD8a (Biolegend; 53-6.7), B220 (Biolegend; RA3-6B2), IgM (Biolegend; RMM-1), CD5 (Biolegend; 53-7.3), and CD11b (Biolegend; M1/70). Cells were analyzed and sorted using a FACS Aria II (BD), as described(Ugarte et al. 2015; Smith-Berdan et al. 2015; Beaudin, Boyer, and Forsberg 2014).

Human B cells: Primary human cells were collected from the blood of a fully consented healthy adult in a study approved by the Institutional Review Board (IRB) at UCSC. For the Tn5Prime analysis, single human B cells were isolated from Peripheral Blood Mononuclear Cells (PBMCs) using negative selection using RosetteSep (StemCell). The resulting B cells were sorted for CD19$^+$ CD27^{high} and CD38^{high}.The following antibodies were used for staining B cells: CD19 (BD Pharmingen; HIB19), CD27 (Biolegend; 0323), and CD38 (Biolegend; HB-7). Cells were sorted using FACS Aria II (BD) and analyzed using FlowJo v10.2 (FlowJo, TreeStar Software, Ashland, OR). For the Smart-seq2 analysis, individual PBMCs were sequenced and B cells for further analysis were identified based on their expression of antibody genes.

Both murine and human single cells were sorted into 96 well plates and directly placed into 4ul of Lysis Buffer - 0.1% Triton X-100, 0.2ul of SuperaseIn (Thermo), 1ul of oligodT primer (IDT), 1ul of dNTP (10mM each)(NEB) - and frozen at -80°C.

RNA-seq library construction and sequencing

2ul of RNA (5 ng) or Single Cell Lysate was reverse transcribed using Smartscribe Reverse Transcriptase (Clontech) in a 10ul reaction including either a Smart-seq2 TSO (Smart-seq2 libraries) or a Nextera A TSO (Tn5Prime libraries) according to manufacturer's instructions for 60min at 42°C (Table S1). The resulting cDNA was treated with 1 ul of 1:10 dilutions of RNAse A (Thermo) and Lambda Exonuclease (NEB) for 30min at 37°C. The treated cDNA was then amplified using KAPA Hifi Readymix 2x (KAPA) and incubated at 95°C for 3 mins, followed by 15 cycles (GM12878) or 27 cycles (single B cells) of (98°C for 20 s, 67°C for 15 s, 72°C for 4 mins), with a final extension at 72°C for 5 mins. For our Tn5Prime method, the cDNA amplification requires both the ISPCR primer and a Nextera A Index primer. For the Smart-seq2 method, the cDNA amplification requires only the ISPCR primer (Table S1). The resulting PCR product was then treated with our Tn5 enzyme (Picelli, Björklund, et al. 2014) custom loaded with either Tn5ME-A/R and Tn5ME-B/R (Smart-seq2) or Tn5ME-B/R adapters only (Tn5Prime). The Tn5 reaction was performed using 5ul of the PCR product, 1ul of the loaded Tn5 enzyme, 10ul of H_2O and 4ul of 5X TAPS-PEG buffer and incubated at 55°C for 5 mins. The Tn5 reaction was then inactivated by the addition of 5ul of 0.2% SDS and 5ul of the product was then nick-translated at 72°C for 6 minutes and further amplified using KAPA Hifi Polymerase (KAPA) using Nextera_Primer_B and Nextera_Primer_A_Universal (Tn5Prime) or Nextera_Primer_A (Smart-seq2) (Table S1) with an incubation of 98°C for 30 s, followed by 13 cycles of (98°C for 10 s, 63°C for 30 s, 72°C for 2 mins) with a final extension at 72°C for 5 minutes. The Tn5 treated PCR product was then size selected using a E-gel 2% EX (Thermo) to a size range of 400-1000bp. GM12878 RNA Smart-seq2 and Tn5Prime libraries were sequenced on an Illumina HiSeq2500 2x150 run, mouse B2 cell Tn5Prime libraries were sequenced on an Illumina MiSeq 2x300 run, human B cell Tn5Prime libraries were sequenced on two Illumina HiSeq3000 2x150 run and human B cell Smart-seq2 libraries were sequenced on a MiSeq 2x75 run.

Sequencing alignment and analysis

Datasets generated from Smart-seq2, Tn5Prime, ENCODE CAGE (GEO accession GSM849368; produced by the lab of Piero Carnici at RIKEN), and ENCODE RNAseq (GEO accession GSM958742; produced by the lab of Barbara Wold at Caltech) (ENCODE Project Consortium 2012) derived from the GM12878 cell line were all trimmed of adapters and low quality bases using trimmomatic (v0.33) (Bolger, Lohse, and Usadel 2014) with a quality cutoff of Q15. Tn5Prime and Smart-seq2 data generated from human single B cells were all trimmed of adapters containing low quality bases using Cutadapt (Martin 2011) and with a quality cutoff of Q15. All paired reads where one or more of the reads contain a post-trimming length of less than 25 bp were filtered out.

Trimmed reads derived from the GM12878 cell line and human single B cells were aligned to the human genome (GRCh38) annotated with Ensembl GRCh38.78 GTF release using STAR (v2.4) (Dobin et al. 2013). Trimmed reads derived from the B2 cells were aligned to the mouse genome (GRCm38) annotated with Ensembl GRCm38.80 GTF release using STAR (v2.4). Expression levels were quantified using featureCounts (v1.4.6-p1) (Liao, Smyth, and Shi 2014) and normalized by total read number resulting in RPM (Reads Per Million).

Peaks for CAGE, Tn5Prime and Smart-seq2 data were called by counting the number of unique fragments which began their forward read alignments (R1 for Tn5Prime) at each position within each chromosome and for each orientation (positive or negative). A peak was called at a position and orientation if at least five alignments begin at that position, the position one nucleotide downstream has fewer alignments beginning at that position, and the position one nucleotide upstream has fewer alignments beginning at that position. For the single cell data, peaks were filtered out unless they appeared in more than one cell. The distance between the Tn5Prime/Smart-seq2 peaks and the nearest CAGE peak was called by inserting the nucleotide coordinates of the CAGE peaks into kd-trees and then performing a nearest neighbor search on them using the Tn5Prime/Smart-seq2 peak coordinates. Each chromosome and orientation had its own kd-tree.

Antibody Assembly

Data generated from our single human B cells were used to identify antibody transcripts. After assigning reads into each cell based upon their cellular index, they were then assembled into transcriptomes using rnaSPAdes (Bankevich et al. 2012) using the single-cell parameters. Putative immunoglobulin transcripts are detected and annotated by running IGBLAST (Ye et al. 2013) against the assembled transcriptome using Human V,D and J segments from the IMGT database (Lefranc et al. 2004). Isotypes are assigned to putative IG transcripts by aligning constant regions to the transcripts with BWA-MEM (H. Li 2013).

Antibody transcripts were filtered using the following process:

1. A table is generated from the SPADES/IGBLAST/BWA pipeline listing each putative IG transcript for each cell in which each row represents one assembled antibody transcript and contains information indicating which cell it came from, overall abundance (as determined by BWA), the CDR3 sequence and the type IGH (Heavy) ,IGK (Kappa) ,IGL (Light) as well as the inferred segments used during VDJ recombination.

2. The transcripts are then clustered by CDR3 sequencing similarity using a single-linkage clustering algorithm Based on the Levenshtein distance where two clusters of transcripts are merged when at least one transcript CDR3 has a Levenshtein distance of 2 or less with the CDR3 of any transcript in another cluster.

3. Transcripts belonging to the same cluster are merged so that highly similar transcripts belonging to the same cell are combined and their transcript counts are added together. This is done to correct for spurious alternative assemblies produced by SPADES within each cell's assembled transcriptome.

4. A list is then generated for each transcript of the cells in which they appear. The lists are then sorted by the transcript abundance within each cell.

6. Each entry in the list are marked by its relative abundance. If the number of reads aligned to the transcript in a cell is less than 10% of the largest number of reads aligned to that transcript within any cell, it is marked as being a potential contaminant.

7. For each type of immunoglobulin transcript (i.e. IGH,IGK,IGL) found within each cell, the largest unique (non-contaminant) transcript (i.e. only found in that cell) is chosen. If a unique transcript cannot be found, then the most highly expressed immunoglobulin transcript is selected.

8. If both a IGK and IGL are present within a cell, the unique transcript is selected. If both are unique or non-unique then the most highly expressed transcript is selected unless either transcript has an abundance of at least 10% of the other.

9. After this elimination process, most cells should have a single heavy chain and light chain.

Visualization

All data visualization was done using Python/Numpy/Scipy/Matplotlib(Hunter 2007; Oliphant 2007; van der Walt, Colbert, and Varoquaux 2011; Jones, Oliphant, and Peterson 2001--). Schematics were drawn in Inkscape (https://inkscape.org/en/).

Data and Script Access

Raw data has been uploaded to the Sequence Read Archive (SRA) and processed gene expression counts are available as supplementary tables S3 (GM12878 Smart-seq2 and Tn5Prime), S4 (Mouse B2 Cells), S5 (Human CD27^{high} CD38^{high} Tn5Prime), and S6 (Human B cells Smart-seq2). Bioproject accession for the SRA are as follows: PRJNA320873 (GM12878 Smart-seq2 and Tn5Prime), PRJNA320902 (Mouse B2 Cells), and PRJNA415475 (Human CD27^{high} CD38^{high} Tn5Prime) and PRJNA433736 (Human B cells Smart-seq2). A UCSC genome browser track is available at

https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=chkcole&hgS_otherUserSessionName=TN5_Prime_Alignments

The Tn5Prime/Smart-seq2, and CAGE Peak Caller and peak distance calculator are available at https://github.com/chkcole/Peak-Calling. All other Scripts are available upon request.

Supplemental

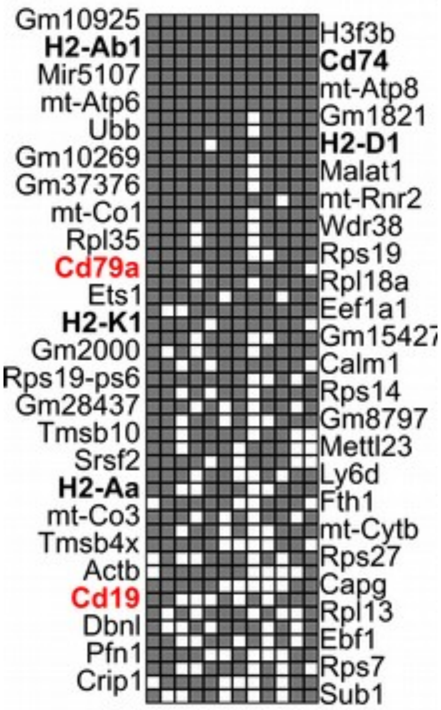


Fig S1.1. Tn5Prime detects B cell specific genes in single B2 cells.

A.) Genes are sorted by the number of B2 cells their transcripts are detected in. Each row represents a gene with its name given on one side. Each column represents are cell. A grey box indicates that gene’s transcript was detected in the respective cell. MHC transcripts are bold. B cell markers in red

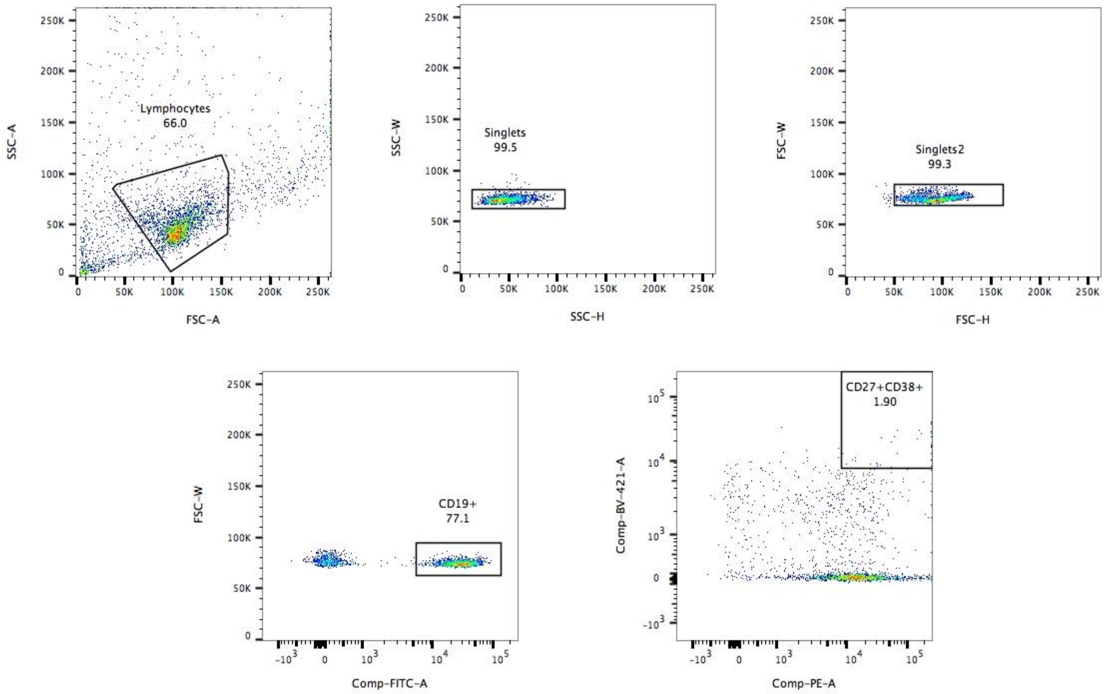


Fig. S1.2. FACS gating strategy for the isolation of CD27^{high} CD38^{high} human B cells

Fluorescence-activated-cell-sorting (FACS) profile of human CD27^{high} and CD38^{high} B cell population. Purified B cells were isolated from PBMCs using a human B cell enrichment method, RosetteSep. Single B cell suspensions were made and stained with anti-CD19-FITC, anti-CD27-BV421 and anti-CD38-PE-A and analyzed using FLOWJO v10.2. Doublets were excluded by SSC-W x SSC-H and FSC-W x FSC-H. Dead cells were excluded using Propidium Iodide staining, only cells which were PI$^-$ were kept for analysis.

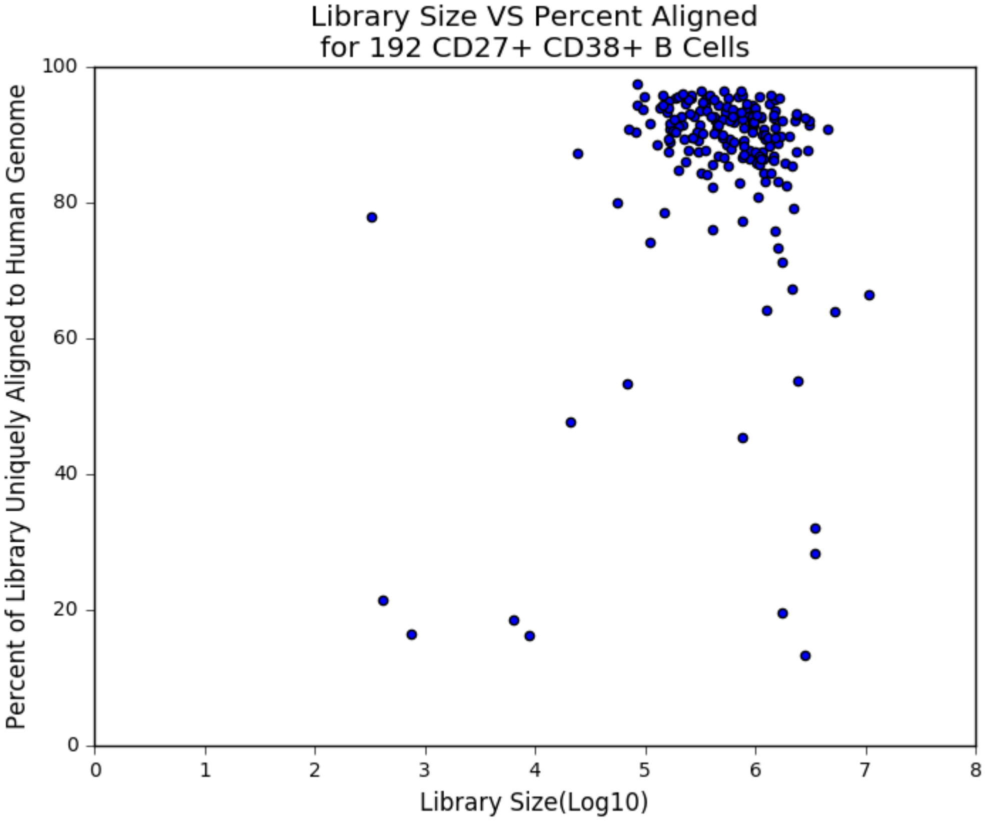


Fig S1.3. Quality Metrics Determined By Alignment Percentage and Read number

A scatter plot of the library size of 192 single CD27+ CD38+ B cells versus their percent alignment. Cells outside more than three median absolute deviations from the median for percent alignment, Mitochondrial transcript percentage, or number of detected genes were marked as outliers and eliminated prior to normalization of transcript counts.

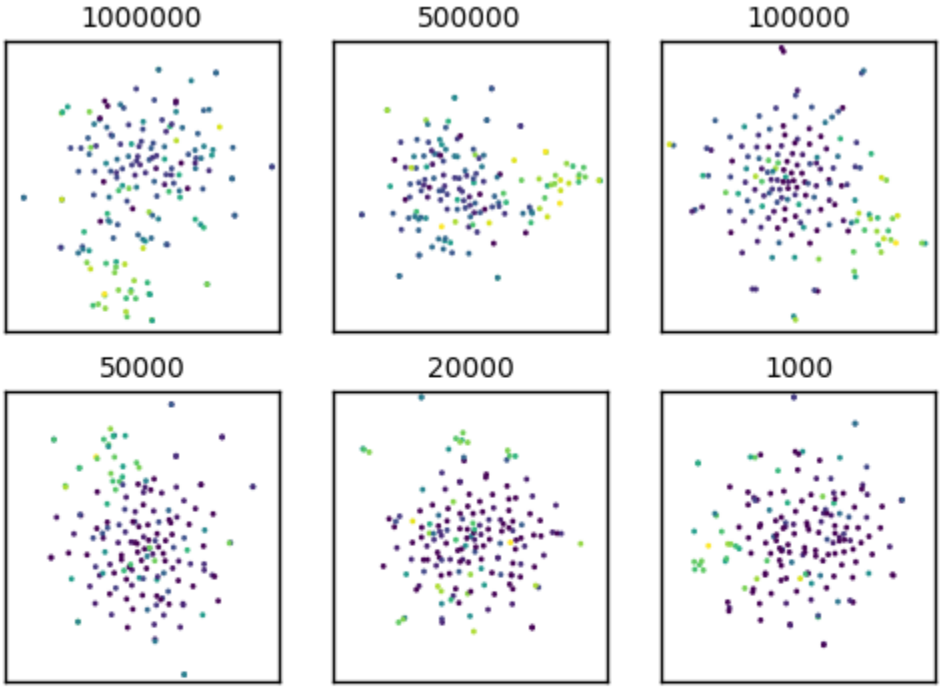


Fig. S1.4. tSNE on subsampled data

In order to test the robustness of the embedding, sampling was performed on each cell's data to a uniform level of depth and t-SNE performed. The embedding was plotted, and cells were colored by the abundance of J-chain. These results indicate that discrimination between the two populations can still be achieved at 100,000 reads, one tenth of the average sequencing depth achieved for each cell.

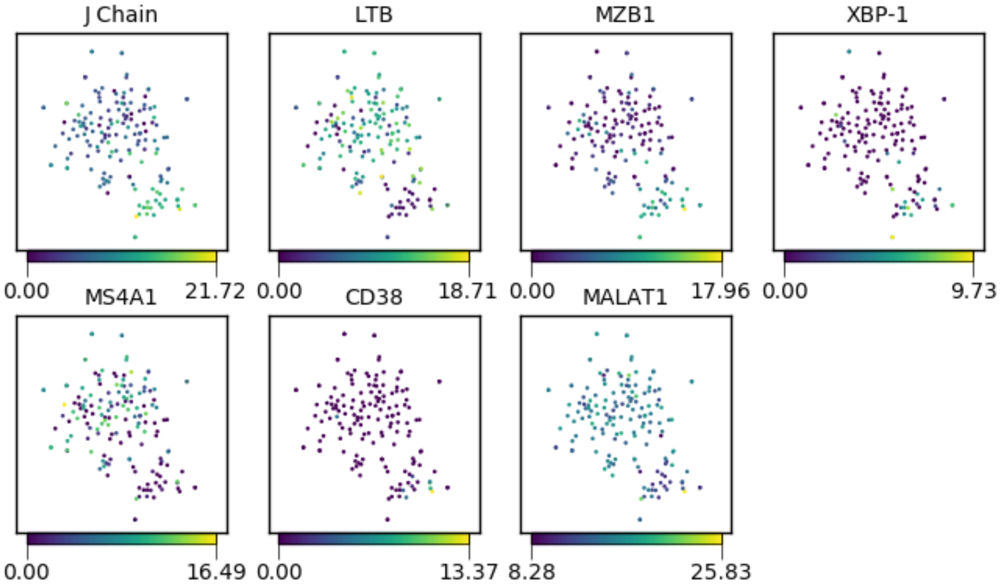


Fig. S1.5. Clustering of CD27+ CD38+ B cells

Cells were plotted in two dimensions using t-SNE based on their normalized transcript counts and colored by the normalized transcript counts of several genes of interest.

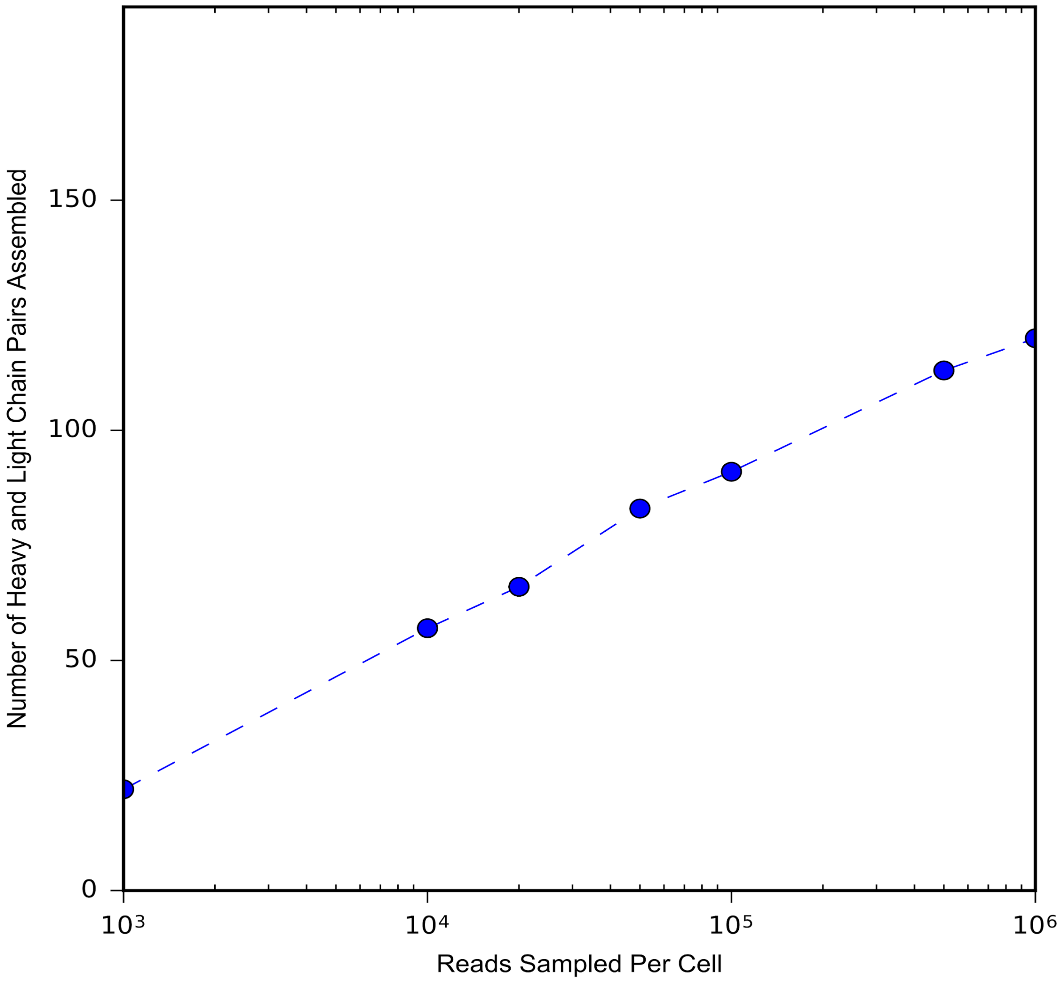


Fig S1.6. Heavy and Chain Pairings Determined by Read Depth

Reads were subsampled in increments of: 1000, 10000, 50000, 100000, 500000 and 1000000 reads per cell to determine read depth necessary for assigning proper heavy and light chain transcripts together. Plot shows on the x-axis reads subsampled per cell and y-axis shows number of light/heavy chain pairs assembled.

RT

<u>Primer</u>

Oligo-dT-smartseq2	/5Me-isodC/AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

<u>TSOs - Tn5Prime</u>

GM12878

TSO_Nextera_Index1	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG GCCTAAGCCATC rGrGrG

B2 cells

TSO_Nextera_Index2	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG GATCTG rGrGrG

Human CD27^{high};CD38^{high}

TSO1_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG rUGA ArU rUC TGGTrGrGrG
TSO2_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG ACrU CrU GrU TGGTrGrGrG
TSO3_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG CrUC rUG rUA TGGTrGrGrG
TSO4_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG rUAG rUA CrU TGGTrGrGrG
TSO5_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG GGrU CrU rUG TGGTrGrGrG
TSO6_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG ArUA GrU ArU TGGTrGrGrG
TSO7_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG rUCC rUA rUC TGGTrGrGrG
TSO8_Nextera	TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG CArU rUC GrU TGGTrGrGrG

<u>TSO - Smartseq2</u>

TSO_Smartseq2	AAGCAGTGGTATCAACGCAGAGTACATrGrGrG

Primers for amplifying cDNA

ISPCR	AAGCAGTGGTATCAACGCAGAGT

Nextera_Primer_A	AATGATACGGCGACCACCGAGATCTACAC [8bp i5 index] TCGTCGGCAGCGTCAGATG

***Tn5 Oligos**

Tn5ME-R	[phos]CTGTCTCTTATACACATCT
Tn5ME-A	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Tn5ME-B	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Primers for amplifying Tn5 Product

Nextera_Primer_B	CAAGCAGAAGACGGCATACGAGAT [8bp i7 index] GTCTCGTGGGCTCGGAGATGTGTAT
Nextera_Primer_A_Universal	AATGATACGGCGACCACCGAGATCTACAC

*Note: Tn5ME-A primer is used for SmartSeq2 protocol.

Table S1.1 Oligos used in the Tn5 Prime Manuscript
All oligos are shown 5’->3’ and were ordered from Integrated DNA Technologies (IDT). Lower case ‘r’ indicates RNA bases. Spaces in sequences are for visual emphasis only.

Chapter 2

Nanopore Long-Read RNAseq Reveals Widespread Transcriptional Variation Among the Surface Receptors of Individual B cells

[This chapter has been adapted from publication, **Nanopore Long-Read RNAseq Reveals Widespread Transcriptional Variation Among the Surface Receptors of Individual B cells**, (Byrne et al., 2017) Nature Communications]

Ashley Byrne[1,2+], Anna E. Beaudin[1,4], Hugh E. Olsen[1,2], Miten Jain[1,2], Charles Cole[1,2],Theron Palmer[1], Rebecca M. DuBois[1], E. Camilla Forsberg[1,3], Mark Akeson[1,2], Christopher Vollmers[1,2,*]

1.) Department of Biomolecular Engineering, University of California–Santa Cruz, Santa Cruz, CA, USA

2.) UC Santa Cruz Genomics Institute, Santa Cruz, California, USA.

3.) Institute for the Biology of Stem Cells, Department of Biomolecular Engineering, University of California–Santa Cruz, Santa Cruz, CA, USA

4) Current Address: Department of Molecular and Cell Biology, School of Natural Sciences, University of California-Merced, Merced, CA, USA

+ Author contribution was data generation, experimental design, data analysis and writing the manuscript.
*Correspondence should be addressed to: C.V. (vollmers@ucsc.edu)

Abstract

Understanding of gene regulation and function requires a genome-wide method capable of capturing both gene expression levels and isoform diversity at the single cell level. Short-read RNAseq, while the current standard for gene expression quantification, is limited in its ability to resolve complex isoforms because it cannot sequence full-length cDNA copies of RNA molecules. Here, we investigated whether full-length RNAseq can be accomplished using the long-read single-molecule Oxford Nanopore MinION sequencing technology (full-length cDNA RNAseq) would be able to identify and quantify complex isoforms without sacrificing accurate gene expression quantification. After successfully benchmarking our experimental and computational approaches on a mixture of synthetic transcripts, we analyzed individual murine B1a cells using a new cellular indexing strategy. Using our computational approaches, we identified thousands of novel, unannotated transcription start and end sites, as well as hundreds of alternative splicing events in these B1a cells. We also identified hundreds of genes expressed across the B1a cells that displayed multiple complex isoforms, including several B cell specific surface receptors and the antibody heavy chain (IGH) locus. Our results show that not only can we identify complex isoforms, but also quantify their expression, at the single cell level.

Introduction

Over the last decade, RNAseq has vastly increased our knowledge of eukaryotic gene expression and the unique transcript isoform signatures that differentiate developmental stages, organs, and single cells (A. R. Wu et al. 2014; Treutlein, Brownfield, et al. 2014; Shalek et al. 2013; Welch, Hu, and Prins 2016). Proteins that arise from transcript isoforms of a single gene can vary in their biological properties including stability, intracellular localization, enzymatic activity, and post-translational modifications (Stamm et al. 2005). Transcript isoforms are the product of alternative transcription start sites (TSSs), transcription end sites (TESs), and alternative splicing events that include alternative splice sites, intron retention, and exon skipping (Sugnet et al. 2004). It has been predicted that a large fraction of human genes are alternatively spliced (Modrek and Lee 2002; E. T. Wang et al. 2008). Although alternative splicing enables increased transcriptome diversity, aberrations in splicing have been implicated in several human diseases, including cancer. In fact, it has been observed that 15% of point mutations are associated with splicing defects, resulting in human genetic disorders (Krawczak, Reiss, and Cooper 1992) and somatic mutations within 12 different cancer types (Brooks et al. 2014).

Consequently, it is important to determine the true transcriptional diversity of cells. This requires that gene expression is analyzed not only at the gene-level but also at the isoform-level. However, current short-read RNAseq methods are inherently limited in their ability to identify complex transcript isoforms, as they cannot sequence full-length transcripts. Instead, transcripts are fragmented for sequencing, resulting in short individual reads that fail to span the entirety of the transcript. Computational tools can be used to assemble full-length transcripts from these reads, but different assembly algorithms can result in conflicting outcomes and varying overall assembly quality (Salzberg et al. 2012).

To offset this limitation of short-read RNAseq, studies have successfully used both single-molecule long-read PacBio and synthetic long-read MOLECULO methodologies (Tilgner et al. 2015; Sharon et al. 2013; Treutlein, Gokce, et al. 2014; Vollmers et al. 2015) to sequence full-length cDNA. However, PacBio technology has a strong bias toward shorter fragments necessitating the separation of cDNA by length before library preparation, which complicates sample preparation and analysis (Bulletin, n.d.). Furthermore, MOLECULO depends on the assembly of short Illumina reads that suffer from the biases inherent in Illumina data and relies on the separation of individual transcript molecules into distinct wells. This complicates quantification as well as the analysis of highly abundant or similar isoforms. Recently, the Oxford Nanopore Technologies (ONT) MinION has been used to analyze full-length cDNA samples derived from both defined synthetic RNA molecules as well as RNA from tissue culture cells (Oikonomopoulos et al. 2016).

With the exception of a single study using single cell RNA-seq to focus its analysis on a single gene locus using PacBio technology (Macaulay et al. 2015), these long-read technologies have been used exclusively to evaluate transcriptome diversity across bulk cell populations. However, recent studies have highlighted that cells found within seemingly homogeneous populations can differ in gene expression (Graf and Stadtfeld 2008),(Irish, Kotecha, and Nolan 2006),(Warren et al. 2006). Understanding heterogeneity within cell populations has shown promise across multiple disciplines including developmental biology, neurobiology, cancer and immunology. Single cell approaches can help illuminate biological questions regarding cell function, development and dysfunction. Knowing the exact state of the cell can help determine its fate or reflect changes with response to stimuli or drug treatment, as well as its ability to neutralize a pathogen, respectively. Cell-to-cell heterogeneity (Shalek et al. 2013) makes immune cells a fascinating target for in-depth analysis of transcriptional diversity. Current approaches that measure RNA transcripts within single cells rely on short-read RNA-seq, single molecule RNA-fluorescence in-situ Hybridization (SM-RNA FISH), or single-cell RT-qPCR (Cornelison and Wold 1997; Raj et al. 2008),(Tang, Lao, and Surani 2011; Femino et al. 1998). These current methods can either be applied to a few genes or are under the same constraints of short-read RNA-seq, which we described earlier. Ultimately, these approaches are unable to identify and quantify complex isoforms on a transcriptome-wide level.

To make it possible to identify and quantify complex isoforms on a transcriptome-wide single cell level, we have developed a nanopore sequencing approach for the analysis of full-length cDNA in single cells. The Oxford Nanopore Technologies (ONT) MinION sequencer is a portable device that is based on single molecule sequencing technology that provides reads of unprecedented length by performing voltage driven molecule translocations through small nanosensors (Cherf et al. 2012). Although the MinION platform has been most useful for interrogating viral and bacterial genomes, recently it has been applied for analyzing cDNA in both targeted as well as genome-wide approaches (Hargreaves and Mulley 2015; Kilianski et al. 2015; Jain et al. 2015; Bolisetty, Rajadinakaran, and Graveley 2015; Oikonomopoulos et al. 2016). Taking advantage of its unprecedented read length, we wanted to interrogate single-cell transcriptomes of mouse B1a cells by sequencing full-length cDNA molecules using the ONT MinION sequencer.

We implemented an integrated informatics pipeline for gene-level and transcript isoform-level expression quantification to overcome the sequencing accuracy limitations of the ONT MinION. To identify transcript isoforms, this pipeline predicted transcription start and end sites, as well as splice sites and their alternative usage. After benchmarking the ONT RNAseq approach on a complex mixture of synthetic transcripts, we sequenced seven individual mouse B1a cells and showed that we could accurately quantify gene expression and identify and quantify novel isoforms at the single-cell level. Our analysis identified differential usage of complex isoforms in over a hundred genes including several surface molecules like CD19, CD20, and IGH, the very receptors defining B cell identity.

Results

Generating and Sequencing Single-Cell RNAseq Libraries

We first investigated the ability of the ONT MinION platform to interrogate transcriptomes at the single-cell level. To test this, we used our ONT RNAseq approach to analyze seven individual mouse B1a cells (Beaudin et al. 2016; Beaudin and Forsberg 2016) and compared it with the standard Illumina RNAseq approach . To this end, we FACS-sorted single B1a cells into individual wells containing lysis buffer and amplified cDNA from each individual cell using the Smartseq2 protocol with modifications (see Methods, Supplementary Table S2.1) (Picelli et al. 2013). The cDNA generated by the Smartseq2 protocol was split and processed in-parallel using the Illumina and ONT library preparation protocols. Sequencing the fragmented cDNA of the seven cells on the HiSeq2500, we generated between 73,086-351,876 150bp Illumina reads per cell. Sequencing the full-length cDNA of the first three cells on individual ONT MinION flow cells using the R7.3 chemistry generated between 17,749-52,696 ONT 2D reads per cell (Supplementary Table S2.2). Taking advantage of the improved MinION throughput using the R9.4 chemistry, we multiplexed the full-length cDNA of the other four cells on a single MinION flow cell and generated between 57,874-128,726 ONT 2D reads per cell. To enable this multiplexing, we introduced custom 60 nucleotide cellular indexes during PCR amplification (see Methods, Supplementary Table S2.1, Fig. 2.1a).

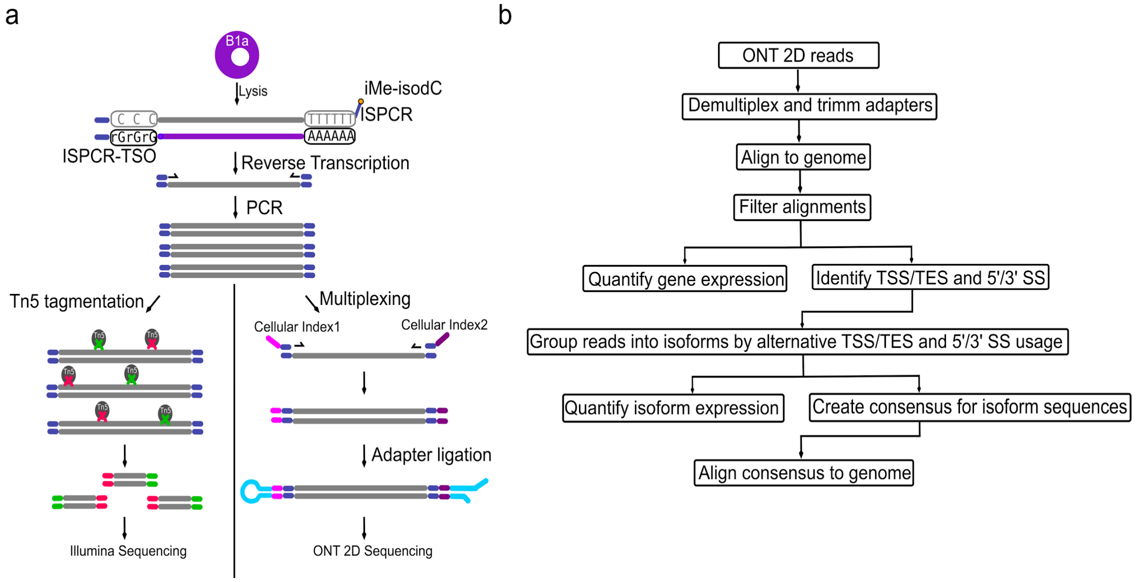


Figure 2.1: Experimental Design and Oxford Nanopore Sequencing read characteristics.

a) Schematic of experimental design. FACS-sorted single B1a cells were lysed. PolyA-RNA was then reverse transcribed, and PCR amplified using template switching. Full-length cDNA was split into two reactions. Half of the reaction was tagmented by Tn5 and sequenced using a Illumina HiSeq2500 sequencer. The other half of the reaction was ligated to ONT adapters and sequenced on an ONT MinION sequencer. b) Schematic of the computational pipeline for ONT 2D read data.

Comparison of Gene Expression Quantification

To assess whether ONT RNAseq is capable of quantifying gene expression, we compared RNAseq data produced with ONT and Illumina, the current benchmark for gene expression quantification. Because standard gene quantification tools (e.g. STAR (Dobin et al. 2013), Cufflinks (Trapnell et al. 2010)) are not compatible with nanopore reads, we aligned the ONT 2D reads using BLAT (Kent 2002) and quantified gene expression using our own algorithm. This algorithm determines how many reads overlap with the exons of a gene to produce a Reads Per Gene per 10K reads (RPG10K) value. As ONT 2D reads are long enough to span the full-length of the transcripts, normalization for gene length was not performed (Fig. 2.1b). Comparing Illumina and ONT RNAseq gene expression quantification for the same cell showed high correlation (Pearson r ≥ 0.84-0.89 for R7.3 and 0.9-0.92 for R9.4), confirming that our ONT RNAseq approach recapitulates Illumina gene expression quantification (Fig. 2.2). Comparing Illumina and ONT RNAseq gene expression quantification across different cells showed low correlation with a Pearson r ≤ 0.45, suggesting that ONT RNAseq can identify cell-to-cell variability (A. R. Wu et al. 2014; Macosko et al. 2015) (Fig. 2.2).

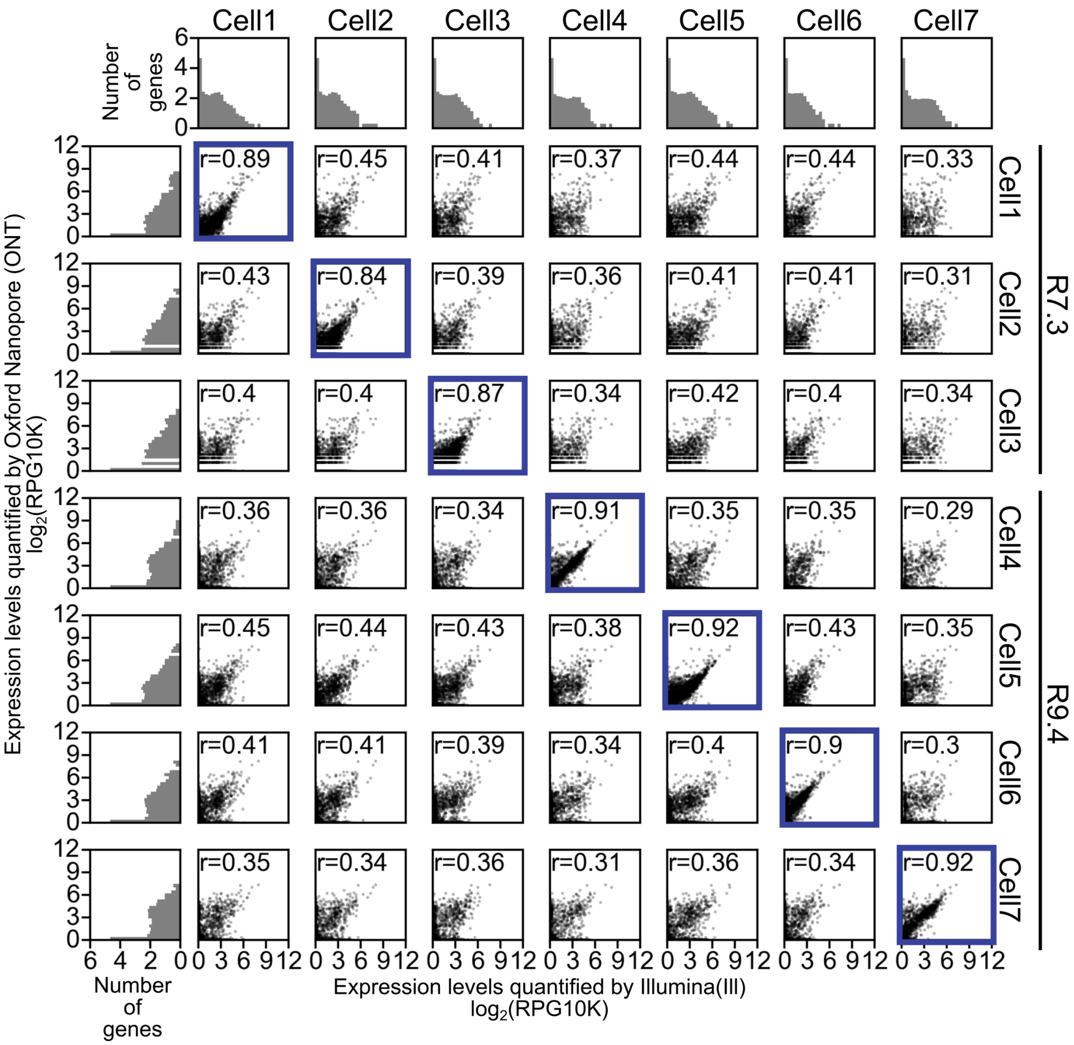


Figure 2.2: ONT RNAseq recapitulates Illumina RNAseq gene expression quantification.

Scatter plot grid at the center of the figure shows gene expression levels for each gene as determined by Illumina RNAseq and ONT RNAseq for the indicated cells. Correlations of transcript expression levels are given as reads per gene per 10,000 reads (RPG10K) across 7 single cells. Pearson r is given for each cell per sequencing method combination with each point representing transcript expression level (x-axes =Illumina and y-axes=ONT). Same cell comparison have a blue border. ONT sequencing chemistry is shown on the right. Histograms found on the left and top of the figure represent number of genes found binned by their expression levels.

These results show that even with the relative low number of reads produced, ONT RNAseq gene-expression quantification largely detects the same genes as Illumina RNAseq (Fig. 2.3a). Furthermore, subsampling ONT and Illumina raw reads showed that, for five of the seven cells analyzed, the detection of expressed genes had reached saturation (Supplementary Fig. 2.2). Unsurprisingly, genes that were detected by either ONT or Illumina RNAseq alone were expressed at lower levels, indicating that these genes were expressed at levels close to the detection limits of both technologies (Fig. 2.3b). We also observed that the genes detected by ONT RNAseq alone were comprised of smaller transcripts (Fig. 2.3c). Additionally, genes that were < 600 bp in length and were detected by both ONT and Illumina RNAseq had relatively lower expression levels in Illumina RNAseq data (Fig. 2.3d). While this is consistent with smaller transcripts being strongly selected against in the Tn5 based Illumina library prep, we couldn't exclude that ONT RNAseq might have a bias towards shorter transcripts. To exclude this possibility, we chose to analyze a mix of synthetic transcripts.

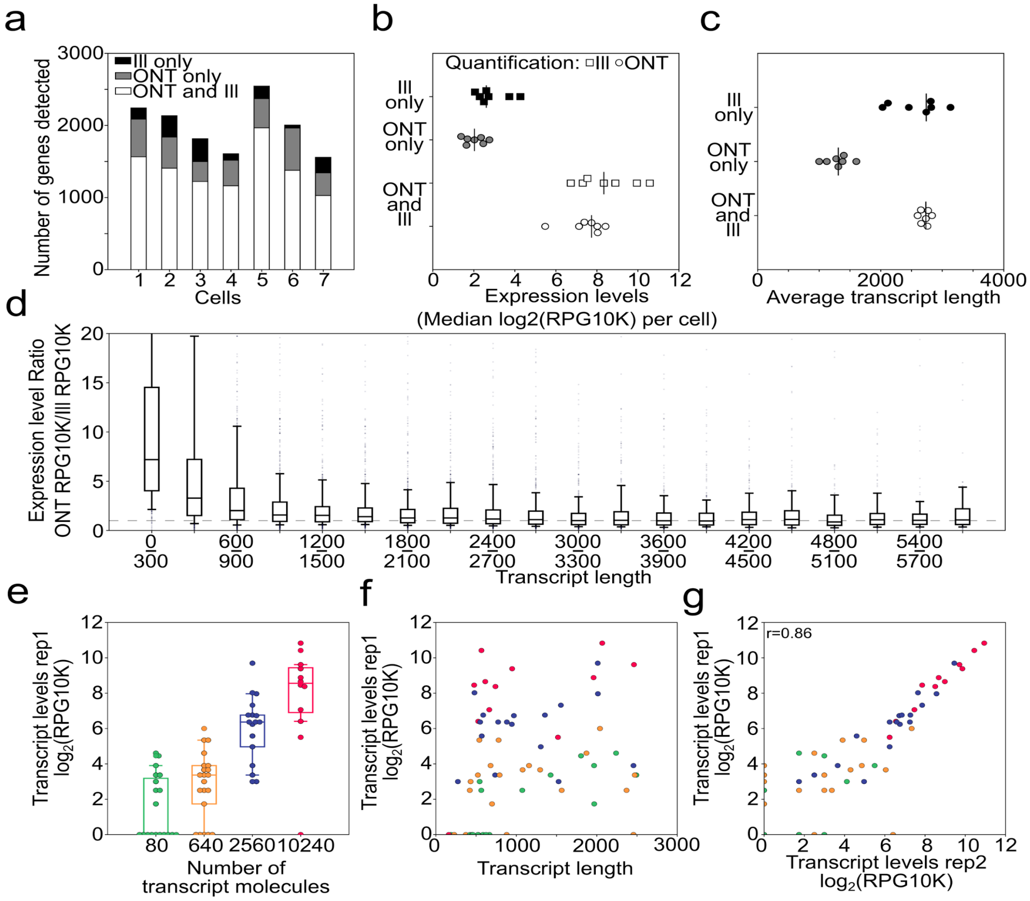


Figure 2.3: Quantifying gene and transcript expression with ONT RNAseq data
a) Stack barplots showing number of genes detected by each cell corresponding to different sequencing technologies (Ill - Illumina, ONT - Oxford Nanopore). b) Median expression levels of genes detected by both or individual technologies. Two expression levels (Ill and ONT) are given for genes detected in both technologies. c) Gene length of genes detected by both or individual technologies. d) Ratio of gene expression levels for genes detected by both technologies. Ratios are binned according to gene length and shown as boxplots with whiskers indicating 10th and 90th percentiles. e) SIRV transcript levels of Replicate 1 (Rep1: 100fg SIRV pool E2) as measured with ONT RNAseq. Transcripts are binned by their starting molecule numbers. f) SIRV transcript levels of Replicate 1 are plotted against transcript length with colors corresponding to groups in e). g) Scatter plot showing correlation of SIRV transcript expression levels of Replicate 1 (Rep1: 100fg SIRV pool E2) and Replicate 2 (Rep2: 100fg SIRV pool E2) , both measured by ONT RNAseq. r-value shown is Pearson-r.

Analysis of synthetic transcript mixtures

To test whether transcript length had an effect on expression levels as measured by ONT RNAseq, we sequenced synthetic Spike-in RNA Variant Control Mixes (SIRVs, Lexogen) of known length, structure and sequence. SIRV transcripts provided in the E2 mix contained 69 transcripts ranging from 191-2528 nt. In the E2 mix 69 transcripts were present in four groups of varying concentrations containing 19, 21, 17 and 12 transcripts in each group, respectively. To test a wide range of possible transcript levels, we amplified (sub-) single cell amounts (i.e. 10fg and 100fg) of the Lexogen SIRV E2 mix in duplicate. This reflected a wide range of possible transcript levels with 8-10,240 molecules of individual SIRV transcripts present before the amplification step.

We quantified the 69 transcripts by aligning the resulting 5367-17915 2D ONT reads directly to the spliced SIRV transcriptome using BLAT and then counting and normalizing the matched ONT 2D reads for each transcript. As expected, when amplifying (sub-) single cell amounts of RNA, we observed transcript drop-out in the lower concentration groups and found that transcript quantification showed variations within each concentration group (Fig. 2.3e, Supplementary Fig. 2.3a). Most importantly, however, quantification was not affected by transcript length, with the exception that transcripts shorter than 500 bp were underrepresented or missed entirely (Fig. 2.3f). Generally, ONT RNAseq quantification agreed with the nominal concentration of the spike-in transcripts and, interestingly, the intra-group variations in transcript quantification were reproducible between replicates (Fig. 2.3g). This intra-group variation might be due to variation in initial transcript levels, systematic amplification bias, or data analysis bias. Overall, the observed underrepresentation of short transcripts in ONT RNAseq and the differences between Illumina and ONT RNAseq quantification are consistent with cDNA molecules below 500 bp in length being selected against during cDNA synthesis and again during the Illumina library preparation using the Tn5 method. Ultimately, analyzing these synthetic transcripts at different concentrations allowed us to exclude the possibility that ONT RNAseq favors shorter transcripts.

Next, we wanted to test whether, in addition to largely unbiased quantification of SIRV transcripts 500-2,500 bp in length, ONT RNAseq reads cover transcripts in their entirety which would make them uniquely suitable to identify and quantify complex isoforms.

Genome annotation and isoform identification with SIRV ONT RNAseq data

The 69 synthetic SIRV transcripts are derived from 7 artificial gene loci that have been modeled after human genes with high isoform diversity, making them suitable for testing ONT RNAseq's capability to capture isoform diversity in a genome annotation independent manner. To this end, we developed algorithms to analyze ONT RNAseq 2D read data to annotate the SIRV gene loci, which in turn could be utilized to further identify and quantify SIRV isoforms. First, we used read alignments to annotate Transcription Start Sites (TSS) and Transcription End Sites (TES), as well as splice sites (SS) of SIRV transcripts in the SIRV gene loci. The annotation of TSS and TES was accomplished by end to end coverage of the entire RNA transcript by complete ONT 2D reads (i.e. reads for which both ISPCR adapters could be identified and trimmed, Supplementary Table 2.2) (Fig. 2.4a-c). Complete ONT 2D reads contained information regarding both TSS, TES in their read alignments.

After combining and aligning ONT 2D reads of all replicates to the artificial SIRV genomic loci (Fig. 2.4c, Supplementary Figure S2.4), we categorized 20 bp bins containing TSS, TESs and splice sites using custom algorithms (see Methods). To avoid the detection of spurious TSS and TES by prematurely terminated read alignments, we required TSS/TES to be at least 60 bp apart. In this manner, we detected 20 TSS and 24 TES that all directly overlapped with an actual TSS and TES and were within 60 bp of 38 (of 57) actual TSSs and 41 (of 59) actual TESs present in the SIRV transcript annotation. Furthermore, we detected 76 (of 89) 5' splice sites and 73 (of 93) 3' splice sites present in the SIRV genome annotation. By analyzing the actual splicing pattern of ONT 2D reads we detected 11 (of 12) alternative 3' splice site combinations and 12 (of 14) alternative 5' splice site combinations as well as 12 (of 12) intron retention events present in the SIRV transcripts.

ONT 2D reads were then sorted into isoform groups based on their TSS/TES and alternative splice site usage. We generated consensus sequences of these groups using POA (Lee, Grasso, and Sharlow 2002) (Partial Ordered Alignment) and compared these consensus sequences to SIRV transcript sequences using BLAT. All of the 33 consensus sequences we generated matched a SIRV transcript with between

97.8% and 100% identity (BLAT identity score) and in all cases matched its directionality. Of the resulting 33 consensus sequences, 26 matched one of the 29 SIRV transcripts present in the two highest abundance groups (Fig. 2.4c,d , Supplementary Fig. 2.4). The other 7 consensus sequences matched one of the 40 SIRV transcripts in the two low abundance groups. While this approach did not succeed in consistently identifying lower abundance isoforms, the consensus isoform sequences detected were very accurate. We also observed high correlation between quantification determined by sorting the reads into their isoform groups and quantification derived from directly aligning reads to the transcriptome (Fig. 2.4e) This means that in addition to identifying sequence, structure, and directionality of complex isoforms, we can also accurately quantify them in a genome annotation independent manner. As a result, we were encouraged to apply this pipeline to our single cell data.

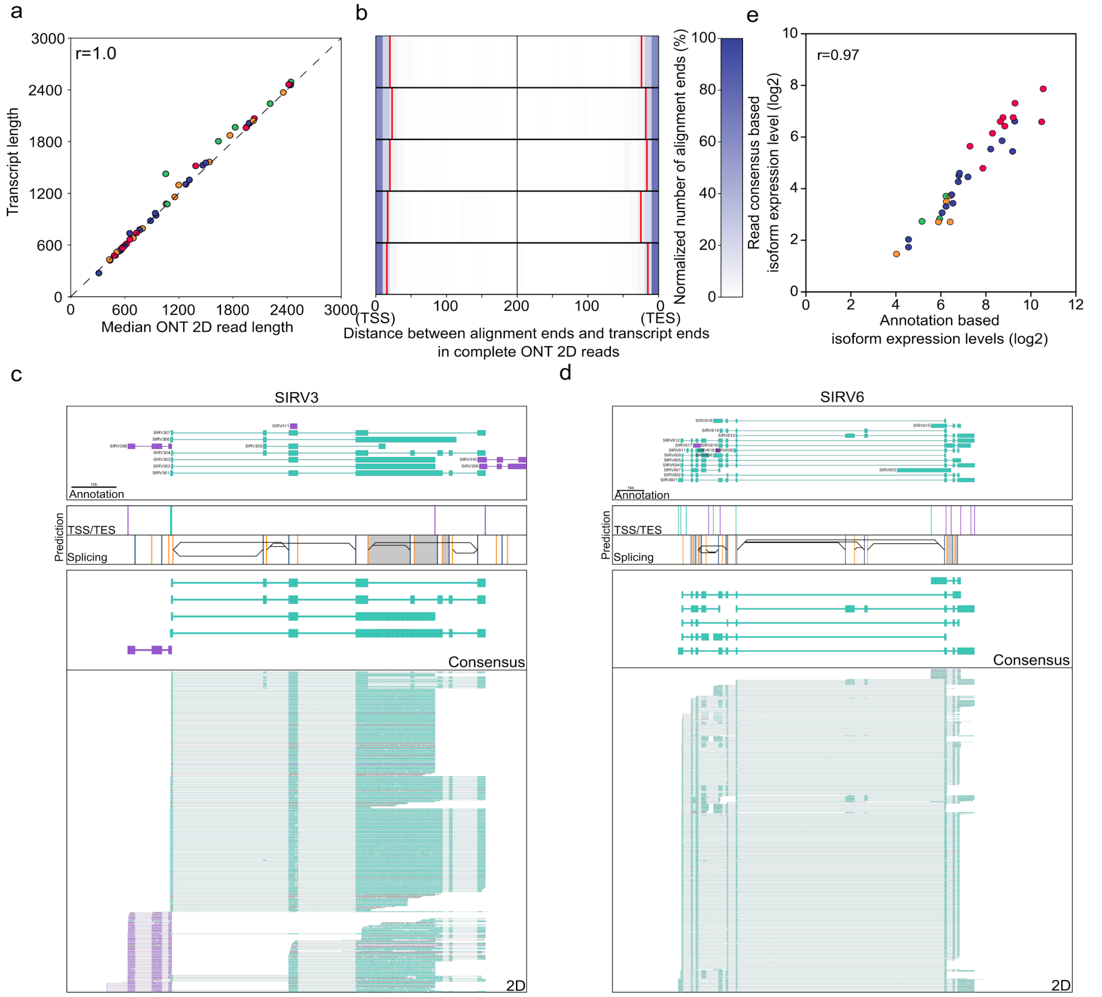


Figure 2.4: Identifying and quantifying isoforms in SIRV E2 mixture
a) Scatter plot shows correlation between ONT 2D reads and SIRV transcripts they align to. Pearson r is shown. Coloring same as Figure 3e-g b) Distance between read alignment ends and transcript ends are shown as heatmap, color indicating the normalized alignment numbers. 90% of read alignments terminated outside the red lines c-d) Genome Browser view of SIRV3(c) and SIRV6(d) gene loci. Top box contains transcript annotations, second and third box contain TSS (Teal) /TES (Purple) and splice sites (5'SS: yellow, 3'SS: blue) locations predicted from the read data. Black lines and grey areas in box 3 indicate alternative splicing and intron retention events predicted from read data. Box 4 contains read alignments of isoform consensus reads. Box 5 contains ONT 2D read alignments. Directionality of ONT 2D reads are indicated by color (Teal: 5' to 3', Purple: 3' to 5'). e.) Scatter plot shows correlation between SIRV transcript quantification by aligning to annotated transcripts or annotation-free isoform grouping. Pearson r is shown.

Identification of Transcription Start and End Sites used in individual B1a cells

By analyzing the ONT 2D reads generated from the seven B1a cells we sequenced, we detected 4234 TSSs and 3883 TESs with only 2476 TSSs and 2448 TESs overlapping with the TSSs or TESs present in the Gencode annotation (vM10) (Stanke et al. 2004; Mudge and Harrow 2015) of the mouse genome (Fig. 2.5a). To determine whether the unannotated TSS and TES we detected were artifacts of our experimental and computational pipelines, we determined their Fantom5 (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014) CAGE peak and polyA signal enrichment. Fantom5 CAGE peaks are derived from capturing and sequencing the 5' end of transcripts and should therefore be enriched in TSSs. Indeed, we found that in contrast to TESs (49/3883 or 1.3%), a high percentage of both annotated (2356/2476 or 95%) and unannotated (1052/1799 or 58%) TSSs overlapped with high scoring Fantom5 CAGE peaks (Fig. 2.5b). Conversely, both annotated and unannotated TESs were highly enriched for polyA signals, while TSSs were not (Fig. 2.5c). When we assigned the detected TSSs and TESs to annotated genes, we found that most genes contained exactly one TSS and one TES, as expected. However, 696 genes contained more than one TSS or TES indicating the presence of more than one isoform (Fig. 2.5d). Overall, this suggested that we successfully identified thousands of unannotated TSSs and TESs and hundreds of genes with differential TSS/TES usage by analysing individual cells.

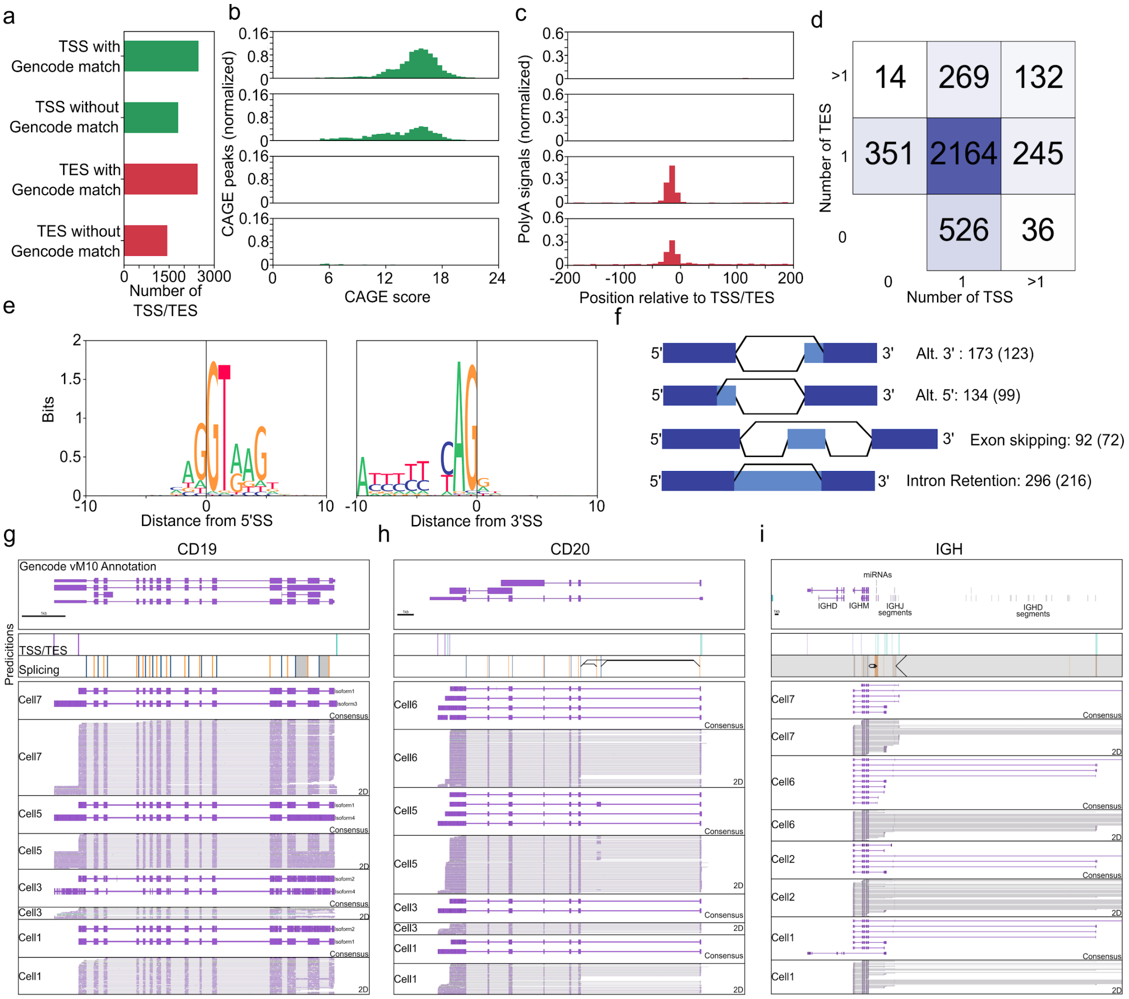


Figure 2.5: Analysis of ONT RNAseq data identifies isoform features in mouse B1A cells

a) TSSs and TESs predicted based on read data were separated into sites with or without GENCODE vM10 annotation matches. b-c) TSSs/TESs with or without GENCODE matches were tested for FANTOM5 CAGE area enrichment (b) and polyA signals (c). d) Overlap of TSSs and TESs with genes. Genes were sorted according to the number of TSSs and TESs they overlapped with. e) Predicted base composition at 5’ and 3’ SS based on read data is shown as sequences logos. f) Schematic for detection and corresponding number of detected alternative splice site combinations. g-i) Genome Browser view of CD19, CD20, and IGH gene loci as shown in Figure 4. ONT 2D reads and consensus sequence alignments are shown for the indicated cells. Splice sites for the highly repetitive IGH locus were not considered for isoform grouping due to the difficulty of aligning reads unambiguously.

Identification of Alternative Splicing Events used in individual B1a cells

In addition to TSSs and TESs, we identified a total of 24,887 5' splice sites and 24,756 3' splice sites. The vast majority of these splice sites were supported by the GENCODE annotation or splice junctions found in Illumina reads. Of the 24,887 5'SS and 24,756 3'SS we identified, 24,298 (97.6%) and 24,220 (97.8%) matched GENCODE annotation, respectively. Of the 589 5'SS and 536 3'SS that did not match GENCODE annotation, 250 (42.4%) and 216 (40.2%) were supported by splice junctions in Illumina reads, respectively. Even if all splice sites that were not supported by GENCODE annotation or Illumina reads were false, which is unlikely, the false discovery rate of our approach would only be 1.3% (659/49,643). Furthermore, while we defined our splice sites as 20 bp bins, we were relatively successful in defining the exact splice site as shown by the base context of the determined splice sites (Fig. 2.5e). By determining alternative splice sites, we found 296 intron retention events, 134 alternative 5' splice sites and 173 alternative 3' splice site combinations. The majority of these events were also observed in Illumina read data, which supported 216 (of 296) intron retention events, 99 (of 134) alternative 5' splice sites, 123 (of 173) alternative 3' splice sites and 72 (of 92) exon skipping events (Fig.2.5f). Alternative events not supported by Illumina read data had significantly lower ONT 2D read counts than those that were supported (Supplementary Table S2.3), indicating they might be closer to the detection limits of both technologies.

Identification of Complex Isoforms

Having established that ONT RNAseq can be used to identify isoform features like TSSs and TESs as well as alternative splicing events, we aimed to identify complex isoforms. We defined genes as expressing complex isoforms if they contained alternative TSS/TES as well as alternative splice sites. We identified 169 genes that expressed complex isoforms. By identifying and quantifying all isoforms we detected of these 169 genes, we found highly significant differential isoform usage between cells in 55 of the genes (Chi2-contigency test, alpha=0.001, holm-sidak multiple-testing correction). These genes with significant differential isoform usage included B cell specific surface receptors CD19 and CD20, the antibody heavy chain locus (IGH) (Fig. 2.5g,h,i), CD37 (Fig. 2.6), as well as CD2 and CD79b, and CD45 (Supplementary Fig. 2.5). We created consensus sequences of the isoforms at these gene loci in each B1a cell and found that across the individual B1a cells, isoforms derived from CD19 showed a combination of alternative TSSs and intron retention events. Isoforms derived from CD20, on the other hand, showed a combination of alternative TESs, as well as an exon skipping event including a previously unannotated exon. The IGH locus was even more complex, with canonical isoforms containing VDJ recombinations and the IGHM constant region exons. In addition, we observed isoforms containing the IGHM constant region exon but originating from 1.) abortive DJ recombinations 2.) I-exon 3.) miRNA loci in the IGHM Switch-region, and 4.) a J-segment. Finally, one isoform in cell 1 originated from the IGHM I-exon but contained the IGHD constant region exons. While IGH isoform diversity has been previously observed and has been known for a long time to be involved in class-switching (K. B. Islam et al. 1994), the ability of ONT RNAseq to sequence full-length cDNA at the single cell level truly highlights and confirms the exceptional transcriptional diversity of the IGH locus.

The ability to sequence entire cDNA molecules from end to end presents an advantage over assembling transcript isoform using Illumina data. While assembling Illumina data using Trinity (Grabherr et al. 2011) is likely to succeed if a gene locus only expresses a single isoform, it appears to struggle with analyzing multiple isoforms of a gene locus that contain multiple distant alternative features. For example, ONT RNAseq identified several distinct isoforms of the CD37 gene across the individual cells analyzed. In most cases, when we assembled the Illumina data from individual cells, Trinity was either unable to form complete contigs or produced contigs that were shown by ONT RNAseq to be misassembled (Fig. 2.6). The CD37 gene and its isoforms therefore highlight the strength of the ONT RNAseq approach to identify the diversity of complex isoforms beyond what is possible with either bulk or short reads technologies.

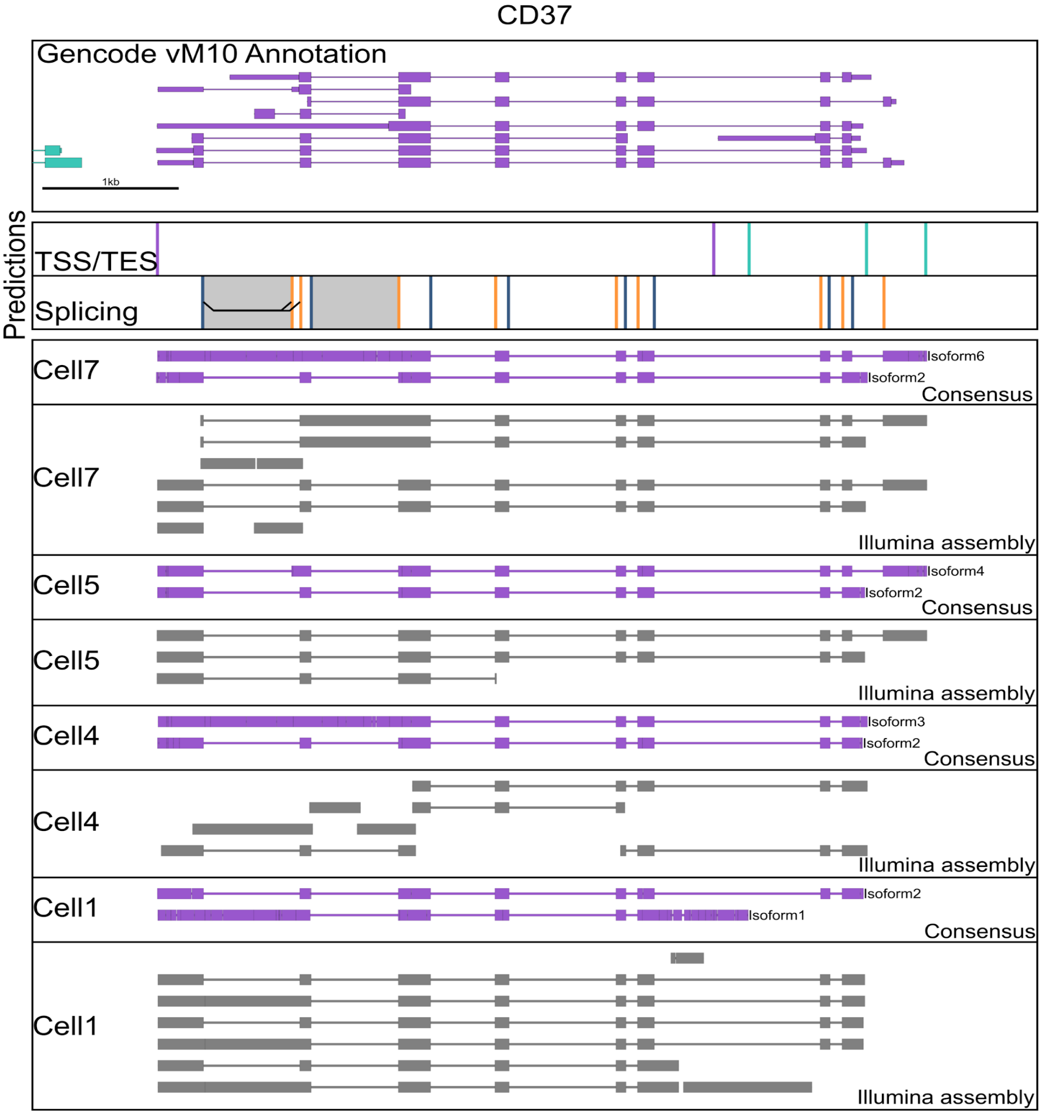


Figure 2.6: Uncovering isoform diversity in B cell surface receptors
Genome Browser view of the CD37.
In addition to isoform consensus derived from ONT 2D reads, contigs assembled from Illumina data using Trinity are shown in grey for the indicated cells.

Discussion

The data we present here shows that RNAseq studies using the Oxford Nanopore Technologies MinION sequencer have the potential to redefine the level of information gathered by a single RNAseq experiment.

By benchmarking our experimental and computational pipelines on ONT MinION data derived from a mix of synthetic transcripts, we showed that our approach identifies the location of transcription start and end sites as well as splice sites in a genome. Furthermore, we have shown that these experimentally determined annotations can then be used to identify and quantify complex isoforms longer than ~500 bp in an otherwise largely transcript length independent manner. It is likely that if we use less stringent size selection methodologies during library preparation, we could capture transcripts < 500 bp as well. Although we were only able to consistently identify the SIRV transcripts found among the high abundance groups, we expect that the less abundant transcripts could be identified using our pipeline by increasing the sequencing depth. Variation in the quantification of transcripts in the SIRV mix indicated that quantification might be improved by using Unique Molecular Identifiers (UMI) (S. Islam et al. 2014) during cDNA amplification. However, UMI length would have to be at least >30bp to be resolved unambiguously with the current error-rate of the ONT MinION. Introducing random nucleotides of this length during priming is likely to create short, unwanted PCR artifacts which would greatly increase the noise of the amplification reaction. Ultimately, until ONT sequencing accuracy improves, the Smart-seq approach employed in this study is currently the best choice for UMI free library generation, as it has been shown by a comparison study to generate the smallest amount of PCR duplicates and the highest transcriptome coverage when comparing low input methodologies (Bhargava et al. 2014).

By focusing on single cells transcriptomes, we demonstrated the capability of sequencing read output and accuracy of ONT MinION sequencer. We showed that ONT RNAseq can not only quantify known genes with a high correlation to Illumina RNAseq but also annotate transcript features, thereby allowing us to identify and quantify complex, never before observed, isoforms. Using ONT RNAseq on only seven B1a cells, we identified thousands of unannotated transcription start and end sites which we then validated using FANTOM5 CAGE data and polyA signals, respectively. Furthermore, we identified 696 genes displaying alternative transcription start and end site usage, and 354 genes with alternative splicing events. Although not all alternative splicing events we detected were supported by single cell Illumina data, the events that weren't supported were of significantly lower coverage, indicating they might be closer to the detection limits in either technology (Supplementary Table 2.3). Combined with the relatively low Illumina sequencing depth per cell in our study, this suggests that larger Illumina depth might aid in the validation of individual events in future studies.

In addition to the identification of individual alternative events, the read length of the ONT MinION sequencer paired with our analysis pipeline enabled us to identify 169 genes expressing complex isoforms containing both alternative TSS/TES and splice sites. Interestingly, among the genes expressing these complex isoforms were surface receptors, including the very surface receptors distinguishing B cells from other immune cells. For example, we found that CD19, CD20 (Ms4a1), IGH, CD45 (B220 or Ptprc), CD2, CD79b, and CD37 were expressed as multiple complex isoforms across the seven B1a cells. This indicates that the diversity of the surface receptors found on B-cells is not fully understood, which could have important implications on all facets of B cell biology. Our data suggest that we are currently only scratching the surface of the true transcriptional diversity of B1a cells. In the future, we aim to use the multiplexing strategy that we have developed to analyze hundreds of individual cells. This will make it possible to truly reconstitute the full transcriptome complexity of B1a and other cell types and will likely lead to discovery of additional subpopulations with distinct functional properties (A. R. Wu et al. 2014). While we currently estimate the cost per cell at ~ $100-200, this is likely to decrease considering the rapidly increasing throughput of the ONT MinION and the soon-to-be-released ONT PromethION sequencer.

Nanopore sequencing is still rapidly maturing and we believe that advancements in sequencing chemistries, nanopore design and analysis algorithms will vastly improve the technology and address the shortcomings of low read numbers and high error rates in the near future. Lower error-rates will, for example, allow us to improve our analysis pipeline further by allowing for the base accurate identification of TSS/TES and splice sites, instead of identifying 20 bp bins for these features. Even with its current limitations, the data and analysis tools we present here demonstrate the potential of ONT RNAseq to revolutionize analysis of transcriptomes. Finally, while the ONT MinION has not quite caught up with the very capable PacBio Sequel long read sequencer, it is only a fraction of its price (~$1,000 vs. $300,000). At this price, any molecular biology lab will be able to perform their own RNAseq experiments on-site, thereby increasing adoption of the single cell RNAseq approach and accelerating research.

Methods

FACS sorting of individual B cells

Mice were maintained in the UCSC vivarium according to IACUC-approved protocols. Single murine Ter119^{-}CD3^{-}CD4^{-}CD8^{-}Gr1^{-}B220^{+} IgM^{+}CD11b^{-}CD5^{+} B1a cells were isolated from wild-type C57Bl/6 mice by lavage and incubated with fluorescently-labeled antibodies prior to sorting (Beaudin et al. 2016). The following antibodies were purchased from Biolegend to stain B-cells: Ter119, CD3 (145-2C11), CD4 (GK1.5), CD8a (53-6.7), B220 (RA3-6B2), Gr1 (RB6-8C5), IgM (RMM-1), CD5 (53-7.3), and CD11b (M1/70). Cells were analyzed and sorted using a FACS Aria II (BD), as described previously (Ugarte et al. 2015), ,(Smith-Berdan et al. 2015),(Beaudin, Boyer, and Forsberg 2014). Single cells were sorted into 96 well plates and directly placed into 4 ul of Lysis Buffer - 0.1% Triton X-100, 0.2 ul of SuperaseIn (Thermo), 1ul of oligodT primer (IDT), 1ul of dNTP (10mM each)(NEB) - and frozen at -80°C.

Smartseq2 cDNA synthesis

Single cell lysate was reverse transcribed using Smartscribe Reverse Transcriptase (Clontech) in a 10 ul reaction including a Smartseq2 (Picelli et al. 2013) TSO (Supplementary Table S1) according to manufacturer's instructions at 42°C. The resulting cDNA was treated with 1 ul of 1:10 dilutions of RNAse A (Thermofisher) and Lambda Exonuclease (NEB) for 30 minutes at 37°C. A PCR amplification step using KAPA Hifi Readymix 2x (KAPA) step was performed incubating at 95°C for 3 mins, followed by 27 cycles of (98°C for 20 s, 67°C for 15 s, 72°C for 4 mins), with a final extension at 72°C for 5 mins.

Illumina Sequencing

The resulting full-length cDNA PCR product was treated with Tn5 enzyme (Picelli, Björklund, et al. 2014) which was loaded with Tn5ME-A/R and Tn5ME-B/R adapters (Supplementary Table S1). The Tn5 treated PCR product was then nick-translated and amplified for 13 cycles with KAPA Hifi Polymerase (KAPA) and Nextera Index Primers (Supplementary Table S1). Libraries were then size selected using a E-gel 2% EX (Thermo-Fisher) to a size range of 400-1000 bp and sequenced on an Illumina HiSeq2500 2x150 run.

Nanopore Sequencing

To achieve the 1ug of DNA needed for the Oxford Nanopore library prep, the full-length cDNA product was split into five aliquots and amplified for 13 cycles with KAPA Hifi Readymix 2X (KAPA) using the ISPCR or multiplex cellular index primers. The following reaction was incubated at 95°C for 3 mins, followed by 13 cycles of (98°C for 20 s, 67°C for 15 s, 72°C for 4 mins) with a final extension at 72°C for 5 mins. The single cDNA or multiplex product was further end-repaired and dA-tailed using NEBNext Ultra End Repair/dA tailing mix (NEB), and adapter ligated using the sequencing adapters provided by ONT (HP Adapter/Adapter Mix). Ligation reaction was performed using Blunt/TA ligase master mix (NEB). Reactions were then enriched using Dynabeads MyOne C1 Streptavidin (Life Technologies) to capture molecules that contain the HP Adapter. Enriched libraries were then mixed with Fuel mix and Running buffer provided by ONT. Single cell libraries were either sequenced solely on one (Cell1 and Cell2) or two (Cell3) separate MinION R7.3 flow cells and ran on the 48 hr 2D protocol. For our multiplexing strategy, single R9.4 flow cells were used (Pool1: Cells4-7, Pool2: Lexogen libraries) and ran on the 48 hr 2D protocol.

Data Analysis

Illumina data

Illumina paired end 150 bp reads in fastq format were quality and adapter trimmed using trimmomatic (v0.33) (Bolger, Lohse, and Usadel 2014). The trimmed reads were aligned using STAR (v2.4) (Dobin et al. 2013) to the mouse genome (Sequence: GRCm38, Annotation: gencode vM10 (Mudge and Harrow 2015)). Illumina reads were assembled for each cell separately using the Trinity (v2.2.0) (Grabherr et al. 2011) set of tools.

ONT data

ONT reads were processed using the Metrichor cloud platform 2D workflow. For R7.3 runs, both reads that passed or failed Metrichor quality cutoffs were retained. For R9.4 runs, reads that failed Metrichor quality cutoffs were discarded as they also failed our alignment criteria. Fast5 files generated by Metrichor were converted into fastq and fasta formats using poretools (v0.5.1) (Loman and Quinlan 2014). For demultiplexing, index-sequences were aligned to the reads using BLAT with parameters: -noHead -stepSize=1 -minScore=20 -minIdentity=20. Reads for which index-sequences could be identified were trimmed and assigned to the respective libraries. Next, for multiplexed and non-multiplexed reads alike, ISPCR adapter sequences were identified and trimmed using Levenshtein distances. Reads for which ISPCR adapters could be identified and trimmed were marked but all reads, trimmed or not, were aligned to the mouse genome (GRCm38) using BLAT(v35x1) (Kent 2002) with parameters: -stepSize=5 -repMatch=2253 -minScore=100 -minIdentity=50 -maxIntron=2000000. Alignments were filtered for a single alignment per read. This filtering process involved three steps: (i) the highest scoring alignment for each read is identified, alignment scores within 2% of each other were treated as ties, (ii) in case of ties the alignment with largest number of gaps is selected (this selects against alignment to unspliced pseudo-genes) and (iii) if the best alignment of a read has a ratio of aligned bases to read bases ≤0.6 the read and its alignment are discarded.

Gene Expression

Gene Expression for ONT and Illumina RNAseq was analyzed using custom scripts. For each gene, the number of reads overlapping with its exons was counted, normalized to total number of aligned reads in a library and reported as Reads Per Gene per 10,000 reads (RPG10K). Genes were counted as expressed if they had a RPG10K value>0. RPG can be calculated as:

$$RPG10K = (total\#\ of\ reads\ aligned\ to\ a\ gene's\ exons \div total\ \#\ of\ aligned\ reads\ in\ sample) \times 10{,}000$$

Transcription start and end site detection

For the detection transcription start and end sites we limited our analysis to reads for which we detected and trimmed ISPCR adapter prior to read alignment. We then identified positions in the genome at which at least 2 alignments of these complete reads ended. We then further restricted our analysis by only considering positions with a median and 75th percentile of the number of clipped (unaligned) read bases between 6-15 and ≤ 20, respectively. This number of clipped (unaligned) bases corresponds to the length of bases contained in ISPCR TSO and oligodT primers that were not trimmed. We then placed a 20 bp bin around these positions to include the highest number of read alignment ends possible.

To filter false positive bins caused by incomplete read alignments in highly expressed genes bins were only considered as containing true TSS/TES if they met the following conditions:

i) The total number of read alignment ends in the bins had to be ≥ 2% of the total number of reads in the next 50 read covered bases. ii) The candidate site had to be at least 60 read covered bases away from the next closest TSS/TES.

By only counting bases covered by read alignments we didn't take non-covered introns into account which would skew our analysis. Next, in order to distinguish TSS and TES bins, we calculated median Levenshtein distances of the unaligned bases at all read alignment ends in a bin to nucleotides present in TSO (ATGG) or the OligodT (TTTT) primer. If the median Levenshtein of a bin to ATGG was ≤2 it was declared a TSS. If the median Levenshtein of a bin to TTTT was ≤2 it was declared a TES.

Transcription start and end site validation

To assess Fantom5(FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014) CAGE scores we downloaded combined CAGE data (mm9.cage_peak_phase1and2combined_coord.bed.gz), converted the data to mm10 coordinates using https://genome.ucsc.edu/cgi-bin/hgLiftOver (Kent et al. 2002) and investigated direct overlap between TSS/TES and CAGE peaks. We considered

TSS/TES and CAGE peaks to be overlapping if they were within 10 bp of each other. To assess polyA enrichment in TSS/TES, we extracted genomic sequences up- and downstream of these sites and looked for identical string matches to “AATAAA” and “ATTAAA”.

Splice Site detection

To identify 20 bp bins as splice sites only ONT 2D reads with a ratio of aligned bases/read bases of > 0.9 were analyzed. We then identified positions in the genome at which at least two read alignments of these reads opened or closed an alignment gap larger than 50 bp.

The 20 bp bins surrounding these positions were considered as containing a splice site if the following conditions are met:

To filter false positive bins caused by spurious read alignment gaps in highly expressed genes bins were only considered as containing true splice sites if they met the following conditions:

i) the number of reads opening or closing an alignment gap in the bin was at least 2% of the total number of reads in the preceding (5’) or subsequent (3’) 40 read covered bases. ii) not closer than 30 bp to another splice site. The directionality of the splice site bin containing either 5’ or 3’ status was based on the direction of the majority of reads containing the splice site.

Alternative Splicing

To detect alternative splice sites, we counted how often 5' and 3' splice sites were spliced together in ONT 2D reads with aligned bases/read bases ratio of > 0.85. A 5'->3' combination had to be present in at least 2 reads to be considered. We scored alternative splice site usage if the same 5' splice site was spliced into two different 3' splice sites or vice versa. To detect intron retentions, we identified areas between 5' and 3' splice sites that were covered to at least 70% by at least one ONT 2D read.

Isoform identification and quantification.

We detected isoform by grouping reads according the TSS/TES and alternative splice sites they contained. ONT read alignment ends found within 60 bp of a TSS and a TES were sorted based on which alternative splice sites it contained. Isoforms that contained at least 1% of all reads at a gene locus were retained. All the reads in these retained isoform groups were used to create consensus reads using POA(Lee, Grasso, and Sharlow 2002). In short, fasta files containing all read sequence are passed to POA which generates a consensus of the reads by creating a multiple sequence alignment of the reads in the form of a partially ordered graph. The program then returns the most heavily-weighted path as the consensus of the reads. The consensus reads are then aligned to genome using BLAT parameters: -stepSize=5 -repMatch=2253 -minScore=10 -minIdentity=10. There was however, one exception regarding the highly complex variable regions derived from the IGH transcripts which were first aligned with IgBlast (Ye et al. 2013) and then with BLAT. IgBlast alignment coordinates were converted to genome coordinates and BLAT and IgBlast portions of the read alignments were merged.

Statistical test and multiple testing correction

We used the ‘chi2_contingency’ function in the scipy.stats (Jones et al., n.d.) package to implement the Chi2 contingency test to detect differential expression of complex isoforms between cells. Multiple testing holm-sidak correction was performed with the ‘statsmodels.sandbox.stats.multicomp’ (Seabold and Perktold, n.d.) package.

Data Visualization

All data analysis and visualization was performed in python (Oliphant 2007) using the numpy/scipy/matplotlib (Jones et al., n.d.; van der Walt, Colbert, and Varoquaux 2011; Hunter 2007) packages.

Data Availability

Illumina and ONT sequencing reads were uploaded to the SRA under accession number SRP082530. All scripts are available upon request or will be available at https://vollmerslab.soe.ucsc.edu/

Supplemental

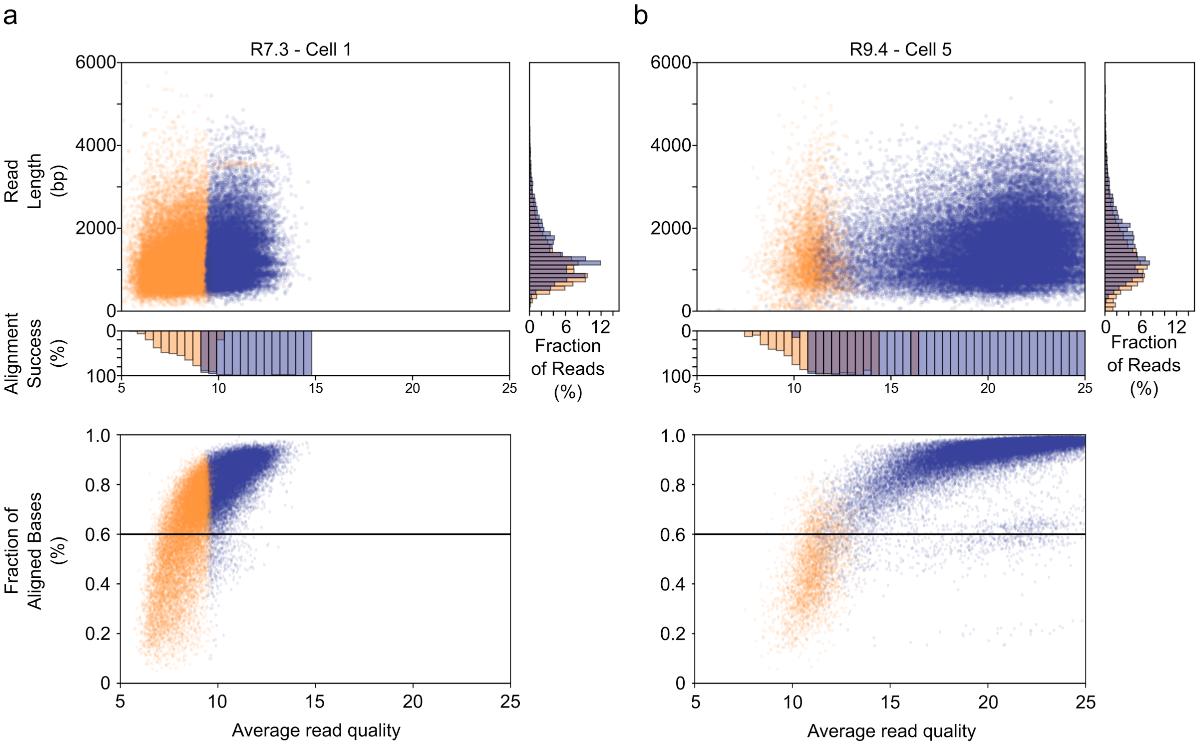


Fig. S2.1: Sequencing run_characteristics
The scatter plots on top shows read length and average sequence quality for ONT 2D reads that passed (blue) or failed (orange) the Metrichor analysis pipeline quality threshold in individual R7.3 (left) and R9.4 (right) sequencing runs. The histograms on the right of the scatter plots show the reads binned by read length using the same colors as the plot in the center to indicate passed (blue) or failed (orange). Using the same color scheme, the histogram at the bottom shows alignment success (percent of reads successfully aligned by BLAT) for reads binned by sequence quality score.
The scatter plot on the bottom shows ONT 2D reads successfully aligned by BLAT. The ratio of aligned/total bases of each read (blue=pass, orange=fail) plotted against average sequence quality score. The alignment quality cut-off of 60% aligned bases is shown as a black line.

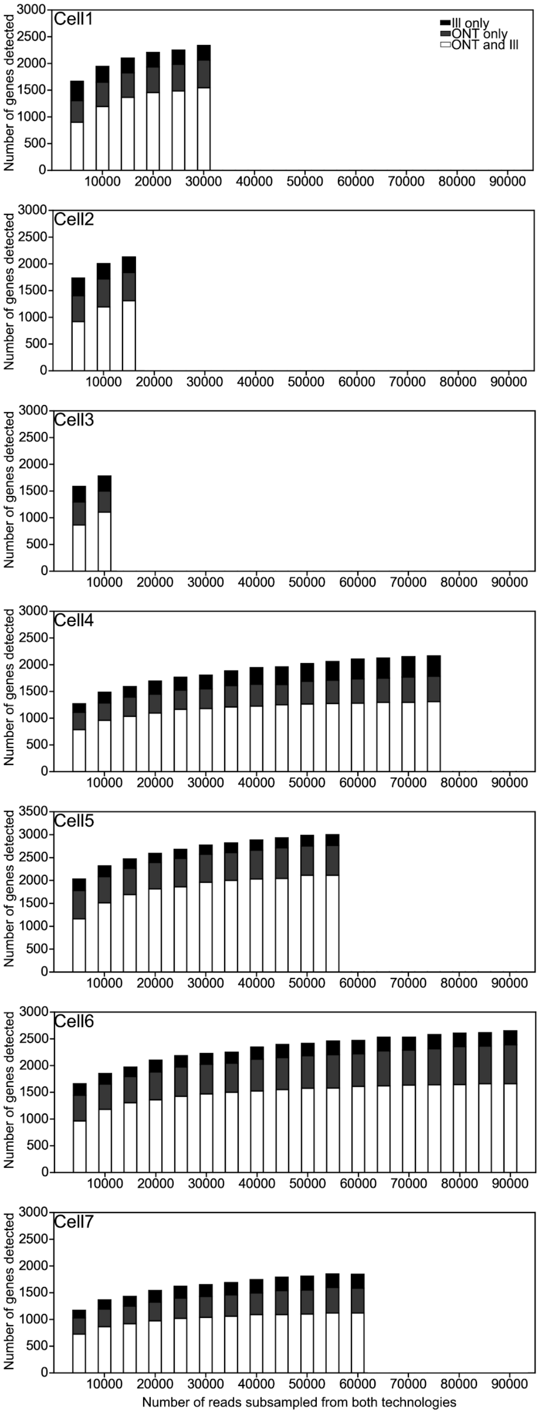


Fig. S2.2: Gene detection with subsampled data
For each cell Illumina and ONT reads were subsampled in 5000 read bins until either the total number of Illumina or ONT reads was reached. The subsampled reads were used to quantify gene expression. Genes with a RPG10K value >0, i.e. with a single mapped read were scored as detected (x-axis = # of genes detected, y-axis= # of reads subsampled). Reads detected in both or either technology are shown in different colored bars.

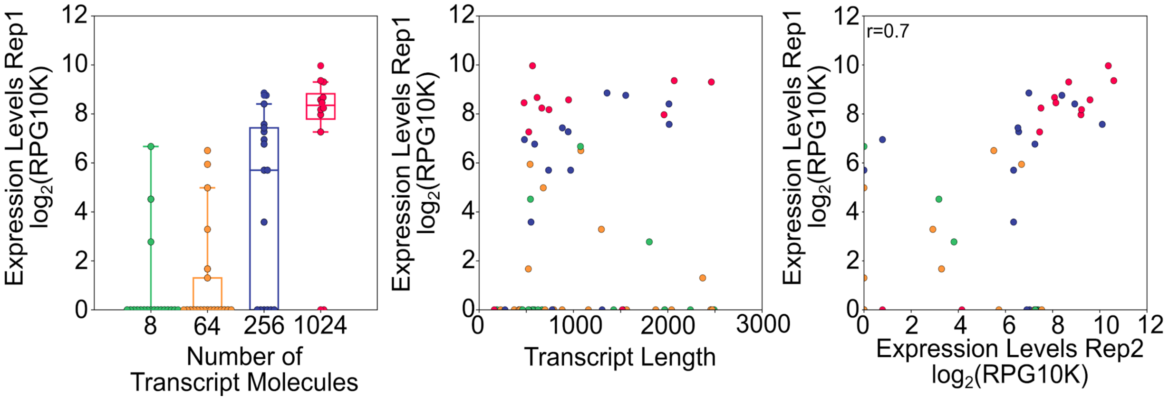


Fig. S2.3: Quantifying SIRV transcripts amplified from 10fg starting material with ONT RNAseq

Left: SIRV transcript levels of Replicate 1 (Rep1: 10fg SIRV pool E2) as measured with ONT RNAseq. Transcripts are binned by their starting molecule numbers.

Middle: SIRV transcript levels of Replicate 1 are plotted against transcript length with colors corresponding to groups in shown on left.

Right: Scatter plot showing correlation of SIRV transcript expression levels of Replicate 1 (Rep1: 10fg SIRV pool E2) and Replicate 2 (Rep2: 10fg SIRV pool E2) , both measured by ONT RNAseq r-value is shown as Pearson-r.

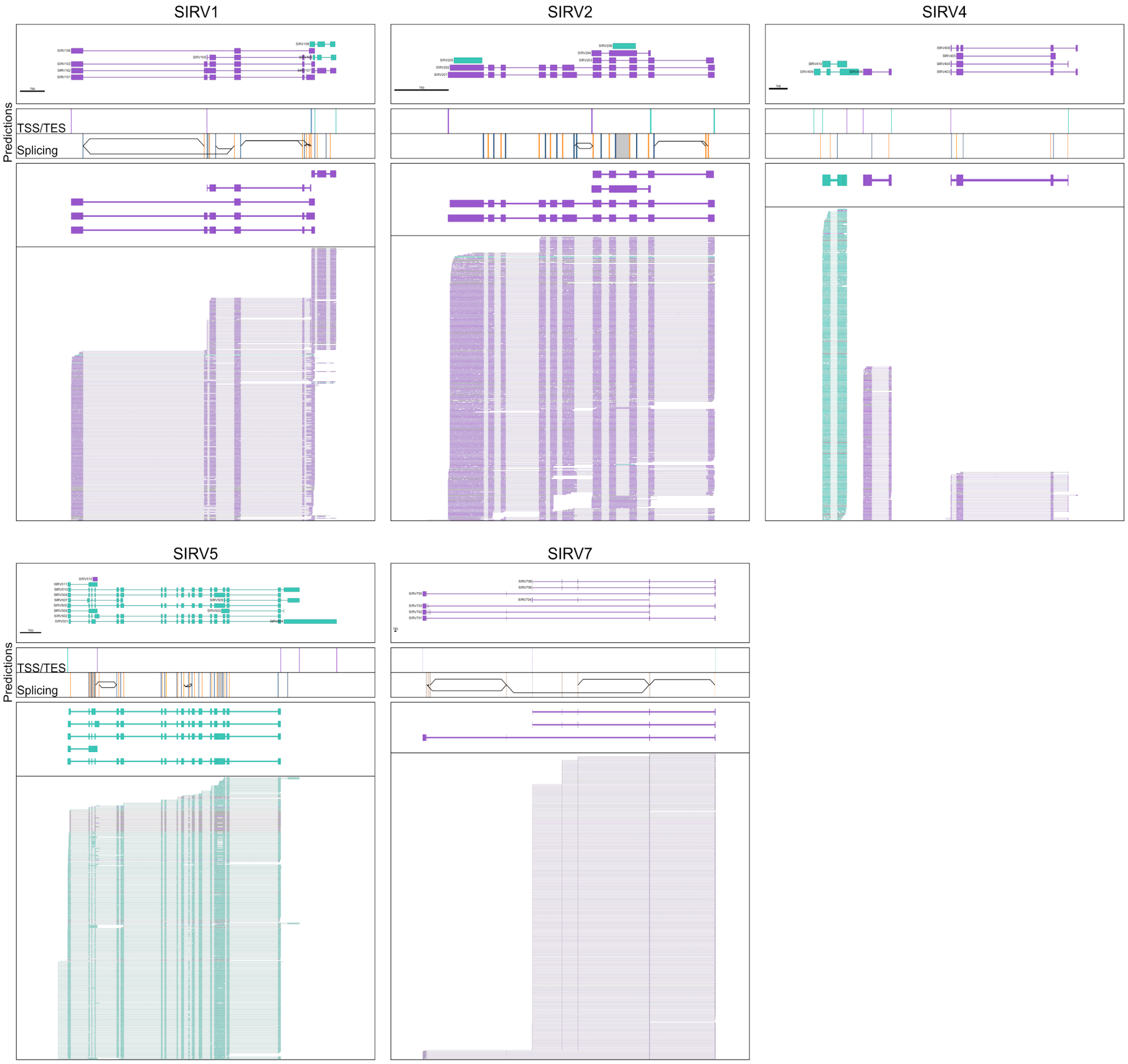


Fig. S2.4: Identifying Transcript isoforms using ONT RNAseq

Genome Browser view of indicated SIRV gene loci. Top box contains transcript annotations, second and third box contain TSS (Teal) /TES (Purple) and splice sites (5'SS: yellow, 3'SS: blue) locations predicted from the read data, respectively. Black lines and grey areas in box 3 indicate alternative splicing and intron retention events predicted from the read data. Box 4 contains read alignments of isoform consensus reads. Box 5 contains ONT 2D read alignments. Direction of transcripts, isoform consensus, and ONT 2D reads are indicated by their color (Teal:5'to3', Purple:3'to5').

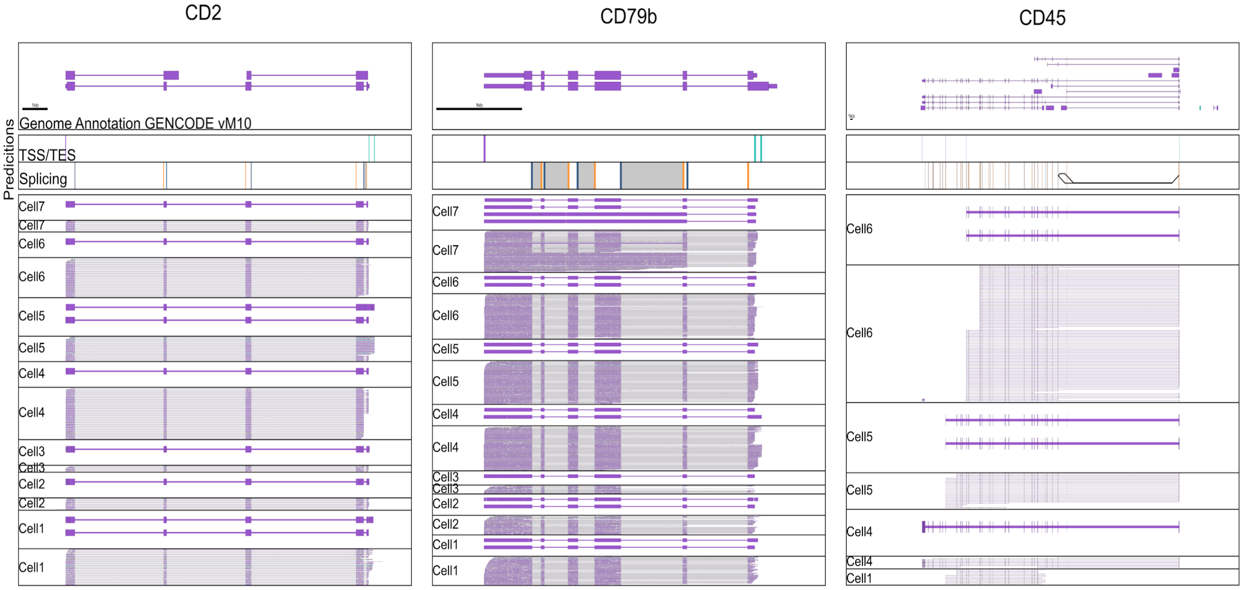
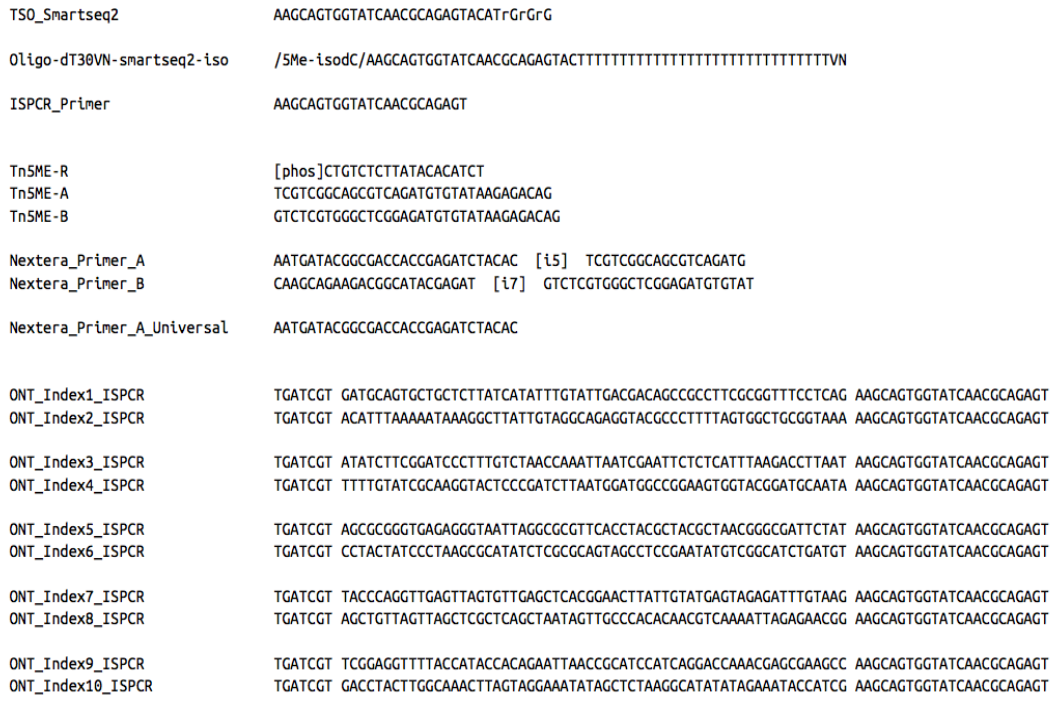


Fig. S2.5: Diverse Isoforms of B cell surface receptors identified using ONT RNAseq

Genome Browser view of the indicated B cell surface receptor gene loci. Top box contains transcript annotations, second and third box contain TSS (Teal) /TES (Purple) and splice site (5'SS: yellow, 3'SS: blue) locations predicted, respectively. Black lines and gray areas in box 3 indicate alternative splicing and intron retention events predicted. Below boxes alternatingly contain read alignments of isoform consensus reads and ONT 2D read alignments. Direction of transcripts, isoform consensus, and ONT 2D reads are indicated by their color (Teal: 5' to 3', Purple: 3' to 5').



TSO_Smartseq2	AAGCAGTGGTATCAACGCAGAGTACATrGrGrG
Oligo-dT30VN-smartseq2-iso	/5Me-isodC/AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
ISPCR_Primer	AAGCAGTGGTATCAACGCAGAGT
Tn5ME-R	[phos]CTGTCTCTTATACACATCT
Tn5ME-A	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Tn5ME-B	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
Nextera_Primer_A	AATGATACGGCGACCACCGAGATCTACAC [i5] TCGTCGGCAGCGTCAGATG
Nextera_Primer_B	CAAGCAGAAGACGGCATACGAGAT [i7] GTCTCGTGGGCTCGGAGATGTGTAT
Nextera_Primer_A_Universal	AATGATACGGCGACCACCGAGATCTACAC
ONT_Index1_ISPCR	TGATCGT GATGCAGTGCTGCTCTTATCATATTTGTATTGACGACAGCCGCCTTCGCGGTTTCCTCAG AAGCAGTGGTATCAACGCAGAGT
ONT_Index2_ISPCR	TGATCGT ACATTTAAAAATAAAGGCTTATTGTAGGCAGAGGTACGCCCTTTTAGTGGCTGCGGTAAA AAGCAGTGGTATCAACGCAGAGT
ONT_Index3_ISPCR	TGATCGT ATATCTTCGGATCCCTTTGTCTAACCAAATTAATCGAATTCTCTCATTTAAGACCTTAAT AAGCAGTGGTATCAACGCAGAGT
ONT_Index4_ISPCR	TGATCGT TTTTGTATCGCAAGGTACTCCCGATCTTAATGGATGGCCGGAAGTGGTACGGATGCAATA AAGCAGTGGTATCAACGCAGAGT
ONT_Index5_ISPCR	TGATCGT AGCGCGGGTGAGAGGGTAATTAGGCGCGTTCACCTACGCTACGCTAACGGGCGATTCTAT AAGCAGTGGTATCAACGCAGAGT
ONT_Index6_ISPCR	TGATCGT CCTACTATCCCTAAGCGCATATCTCGCGCAGTAGCCTCCGAATATGTCGGCATCTGATGT AAGCAGTGGTATCAACGCAGAGT
ONT_Index7_ISPCR	TGATCGT TACCCAGGTTGAGTTAGTGTTGAGCTCACGGAACTTATTGTATGAGTAGAGATTTGTAAG AAGCAGTGGTATCAACGCAGAGT
ONT_Index8_ISPCR	TGATCGT AGCTGTTAGTTAGCTCGCTCAGCTAATAGTTGCCCACACAACGTCAAAATTAGAGAACGG AAGCAGTGGTATCAACGCAGAGT
ONT_Index9_ISPCR	TGATCGT TCGGAGGTTTTACCATACCACAGAATTAACCGCATCCATCAGGACCAAACGAGCGAAGCC AAGCAGTGGTATCAACGCAGAGT
ONT_Index10_ISPCR	TGATCGT GACCTACTTGGCAAACTTAGTAGGAAATATAGCTCTAAGGCATATATAGAAATACCATCG AAGCAGTGGTATCAACGCAGAGT

Table S.2.1 List of Oligos

All oligos in the study are shown above

1) Spaces in sequences are for visual emphasis and do not indicate gaps in the sequences. 2) Nextera_Primer_A and Nextera_Primer_B represent groups of sequences. 3) (i5) and (i7) indicate different illumina indexes

		Illumina Reads		ONT 2D Reads				Percent of pass alignment filter ONT 2D reads		
Technology	Sample	Raw Reads	Aligned	Pass Filter	Fail Filter	Analyzed	Pass Alignment Filter	Complete Reads: Both ISPCR sequences found	1st ISPCR sequence found	2nd ISPCR sequence found
R7.3	Cell1	351876	321251	23957	28739	52696	32556	38.16	71.71	55.91
R7.3	Cell2	185258	157782	8665	13396	22061	13889	36.85	69.00	57.38
R7.3	Cell3	109712	89998	5081	12668	17749	8500	36.27	67.86	56.29
R9.4	Cell4	121586	94237	76721	1936	76721	74604	62.32	76.51	81.29
R9.4	Cell5	174128	158788	53651	4223	53651	52304	66.47	79.11	84.10
R9.4	Cell6	97314	89572	127408	1318	127408	124007	66.50	79.99	83.20
R9.4	Cell7	73086	59738	105056	2432	105056	102412	64.01	76.31	83.89
R9.4	Lexogen_10fg_Rep1	--	--	12725	368	12725	10837	68.11	80.37	84.47
R9.4	Lexogen_10fg_Rep2	--	--	17915	412	17915	15327	70.94	82.10	86.12
R9.4	Lexogen_100fg_Rep1	--	--	5367	526	5367	4770	66.98	79.87	83.63
R9.4	Lexogen_100fg_Rep2	--	--	5745	528	5745	5197	65.25	78.01	83.41

Table S.2.2: Illumina and ONT Read Numbers

Illumina Reads: Numbers indicate individual reads, not read pairs. “Aligned” Reads are reads successfully aligned using STAR. ONT 2D Reads: “Pass Filter” and “Fail Filter” are reads determined by Metrichor software based on quality scores. “Pass Alignment Filter” are reads that were aligned using BLAT and more than 60% of their bases aligned to the genome. Complete reads as mentioned in the manuscript are defined as reads for which both the ISPCR sequences are identified on both ends and trimmed.

	Median read count		p-value (MWU one-sided)
Alternative 5' SS	81	42.5	0.0003
Alternative 3' SS	78	33	0.0019
Intron Retention	15	10.5	0.0412
	yes	no	
	Illumina supported		

Table S2.3: Alternative Splice Site Predictions

Median ONT 2D read count is shown for alternative splice sites. Splice sites are separated into sites with and without illumina read support. p-values are calculated using scipy.stats.mannwhitneyu function with keyword argument alternative = ‘greater’.

Depletion of Hemoglobin Transcripts and Long Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (*Ursus* maritimus)

Ashley Byrne[1,2+], Megan A. Supple[3], Roger Volden[2,4], Kristin L. Laidre[5], Beth Shapiro[3,6], Christopher Vollmers[2,4,*]

[This chapter has been adapted from publication, **Depletion of Hemoglobin Transcripts and Long Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (*Ursus* maritimus)**, (Byrne et. al, 2019) Frontiers of Genetics]

1) Department of Molecular, Cellular, and Developmental Biology, University of California Santa Cruz, CA 95064

2) Genomics Institute, University of California, Santa Cruz, CA 95064

3) Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064

4) Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064

5) Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle, WA 98195

6) Howard Hughes Medical Institute

*** CORRESPONDENCE:** Dr. Christopher Vollmers vollmers@ucsc.edu

+ Author contributions include conception of method, experimental design, generating data, data analysis and writing manuscript

Abstract

Transcriptome studies evaluating whole blood and tissues are often confounded by overrepresentation of highly abundant transcripts. These abundant transcripts are problematic as they compete with and prevent the detection of rare RNA transcripts, obscuring their biological importance. This issue is more pronounced when using long-read sequencing technologies for isoform-level transcriptome analysis, as they have relatively lower throughput compared to short-read sequencers. As a result, long-read based transcriptome analysis is prohibitively expensive for non-model organisms. While there are off-the-shelf kits available for select model organisms capable of depleting highly abundant transcripts for alpha (HBA) and beta (HBB) hemoglobin, they are unsuitable for non-model organisms. To address this, we have adapted the recent CRISPR/Cas9 based depletion method (Depletion of Abundant Sequences by Hybridization) for long-read full-length cDNA sequencing approaches that we call Long-DASH. Using a recombinant Cas9 protein with appropriate guide RNAs, full-length hemoglobin transcripts can be depleted *in-vitro* prior to performing any short- and long-read sequencing library preparations. Using this method, we sequenced depleted full-length cDNA in parallel using both our Oxford Nanopore Technology (ONT) based R2C2 long-read approach, as well as the Illumina short-read based Smart-seq2 approach. To showcase this, we have applied our methods to create an isoform-level transcriptome from whole blood samples derived from three polar bears (*Ursus maritimus*). Using Long-DASH, we succeeded in depleting hemoglobin transcripts and

generated deep Smart-seq2 Illumina datasets and 3.8 million R2C2 full-length cDNA consensus reads. Applying Long-DASH with our isoform identification pipeline, Mandalorion we discovered ~6,000 high-confidence isoforms and a number of novel genes. This indicates there is a high diversity of gene isoforms within *Ursus maritimus* not yet reported. This reproducible and straightforward approach has not only improved the polar bear transcriptome annotations but will serve as the foundation for future efforts to investigate transcriptional dynamics within the 19 polar bear subpopulations around the Arctic.

Introduction

Accurate isoform-level differential expression analysis of transcriptomes is essential for interpreting gene regulation under different biological, environmental or physiological conditions. RNA transcript isoforms – which are often unique among different cell types, tissues, developmental stages, and organisms (Zhang et al. 2016; E. T. Wang et al. 2008; Kalsotra et al. 2008) – are defined by the use of alternative transcription start sites (TSSs), polyA sites, and splice sites. Use of alternative isoforms is highly regulated and thought to contribute to cellular and organismal diversification within higher eukaryotes (Graveley 2001), adaptation and speciation (Harr and Turner 2010; Shi et al. 2012) and can also reflect certain disease states (Busslinger, Moschonas, and Flavell 1981; Andreadis 2005; Ilagan et al. 2015).

To perform this type of differential expression analysis on the isoform-level requires both short- and long-read sequencing technology. Short-read RNA-seq provides the read depth necessary for gene expression quantification, but requires accurate and exhaustive isoform-level transcriptome annotations for its analysis. However, existing transcriptome annotations of non-model organisms are often incomplete or inaccurate (Ungaro et al. 2017) because they cannot rely on labor-intensive efforts like Gencode, which are working to exhaustively annotate the isoform-level transcriptomes of human and mouse. While short-read RNA-seq data can itself be used for transcriptome annotation, it fails at annotating transcriptomes on the isoform-level because it cannot recapitulate full-length transcripts. This inability to define full-length transcripts is due to the fragmentation of RNA, or their cDNA copies, prior to sequencing making it difficult to computationally re-assemble reliably (Grabherr et al. 2011; Pertea et al. 2015; Bankevich et al. 2012). To provide an accurate isoform-level transcriptome annotation for non-model organisms, long-read sequencing technology is required to sequence full-length cDNA molecules.

The ability to perform combined short- and long-read transcriptome analysis on non-model organisms is further complicated by sample availability. In contrast to the organs and tissues of model organisms which can be easily acquired, availability of samples from non-model organisms are often more limited. In rare circumstances sampling can be performed through fat and muscle tissue biopsies (Khudyakov et al. 2017), but the current gold standard still relies on whole blood RNA samples, especially for large non-model organisms (Du et al. 2015). This is particularly true for protected and endangered species (Huang et al. 2016; Hernández-Fernández, Pinzón, and Mariño-Ramírez 2017). While whole blood samples can be easily acquired and provide a wealth of information regarding physiological or disease states in surrounding tissues (Liew et al. 2006), polyadenylated RNA extracted from whole blood can be comprised of >50% hemoglobin transcripts (Shin et al. 2014; Mastrokolias et al. 2012). In any high-throughput sequencing-based assay, these highly abundant transcripts will compete for a limited number of sequencing reads and as a result will be sequenced over and over again without generating any new information. This would waste valuable reads which could otherwise detect less abundant transcripts.

Currently, there is no approach to deplete hemoglobin transcripts from whole blood RNA while enabling downstream analysis of the depleted RNA/cDNA with both short- and long-read sequencing. Commercially available hemoglobin depletion kits – including GLOBINclear (Ambion) or RiboZero (Illumina) – are specifically designed for human samples and rely on hemoglobin RNA pull-down methods (Field et al. 2007). Even if they would succeed in depleting hemoglobin from non-model organism samples, which is far from guaranteed (Choi et al. 2014), these pull-down approaches use harsh conditions and high temperatures during long incubation steps which contribute to RNA fragmentation and introduce unwanted technical variability (Debey et al. 2004). While fragmented RNA is suitable as input into short-read RNA-seq, it is not suitable for long-read full-length cDNA sequencing.

To perform a comprehensive analysis of non-model organism transcriptomes from whole-blood with short- and long-read technologies, we require a new approach that can deplete highly abundant transcripts like hemoglobin from whole-blood samples of a wide range of organisms without fragmenting transcripts. To this end, we chose to adapt the powerful, recently published DASH (Depletion of Abundant Sequences by Hybridization) (Gu et al. 2016) method which utilizes a recombinant Cas9 to perform *in-vitro* depletion using sequence specific sgRNAs. Our adapted method which we will refer to as Long-DASH also takes advantage of the CRISPR/Cas9 system to selectively deplete hemoglobin alpha (HBA) and beta (HBB) transcripts but targets full-length cDNA instead of fragmented short-read Illumina sequencing libraries like regular DASH. By depleting full-length cDNA prior to any library preparation, this allows the user the choice to use any short- or long-read sequencing platform.

As a proof-of-concept we evaluated three hemoglobin-depleted and non-depleted polar bear whole blood transcriptomes using our ONT-based R2C2 (Volden et al. 2018) full-length cDNA sequencing method and an Illumina-based modified Smart-seq2 method. We found that by incorporating Long-DASH, we successfully depleted hemoglobin transcripts without non-specifically affecting the rest of the cDNA pool. Finally, we generated ~3.8 million ONT-based R2C2 consensus reads, dramatically refining the polar bear transcriptome annotations.

RESULTS

Long-DASH depletes hemoglobin transcripts from full-length cDNA

We used a modified Smart-seq2 protocol (Picelli, Faridani, et al. 2014; Cole et al. 2018; Volden et al. 2018) to reverse transcribe and amplify full-length cDNA from 70 ng of whole blood RNA of three polar bears (PB3, PB19, PB21). We then performed a targeted depletion of hemoglobin transcripts by incubating the full-length cDNA with Cas9 protein loaded with 16 guide RNAs (sgRNAs) specific to hemoglobin transcripts – 8 sgRNAs targeting the HBA transcripts and 8 sgRNAs targeting the HBB transcripts. The sgRNAs were selected to deplete hemoglobin transcripts from human and polar bear samples. The sgRNAs were chosen based upon sequence homology between these two species to eventually allow the removal of abundant of hemoglobin transcripts in whole blood from both human and polar bear samples using the same sgRNAs (Field et al. 2007) (Supplementary Fig. S3.1). Using the 16 sgRNA probes we designed should also allow for the depletion of samples of other vertebrates although sequence similarity should be checked before this is attempted.

The depletion process using the Cas9 system should cut the ~700-800 bp transcripts at different sites allowing us to do two things. First, we can re-amplify the sample, thereby only enriching for full-length molecules since the cut cDNA molecules no longer contain two priming sites required for exponential amplification during PCR amplification (Fig. 3.1). Second, we can remove the cut transcripts by performing a SPRI-bead based size selection whereby only transcripts > 500 bp are retained. Indeed, prior to any depletion, we observed very strong bands located at ~700-800 bp in our agarose gels indicating the presence of a substantial amount of HBA and HBB hemoglobin transcripts (Fig. 3.2). After depletion, reamplification, and size selection, the full-length cDNA product was visualized again to reveal the removal of the putative hemoglobin bands (Fig. 3.2). After hemoglobin depletion is confirmed, the cDNA is ready to be converted into ONT- and Illumina-based libraries, with each protocol using the same input cDNA (Fig. 3.1).

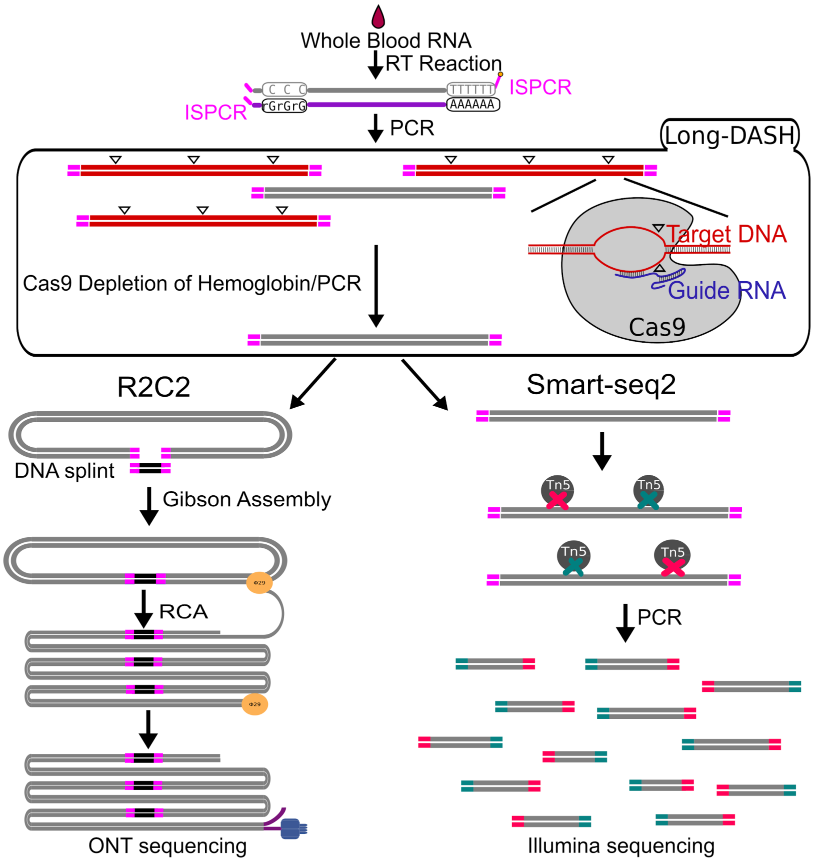


Figure 3.1: Schematic of Long-DASH (Long Read Depletion of Abundant Sequences by Hybridization)

A) Whole Blood RNA is extracted and full-length cDNA is generated with the first half of the Smart-seq2 protocol. The cDNA is then depleted of hemoglobin transcripts using the recombinant S. pyogenes Cas9 protein bound to hemoglobin specific sgRNA which cuts hemoglobin cDNA molecules by introducing double strand breaks (Δ) in a sequence specific manner. The cut molecules can no longer be exponentially amplified with PCR, so a subsequent PCR step is performed to enrich for complete non-hemoglobin cDNA molecules. The resulting hemoglobin depleted cDNA pool is then sequenced using the ONT-based R2C2 library prep and the Illumina-based Smart-seq2 library prep.

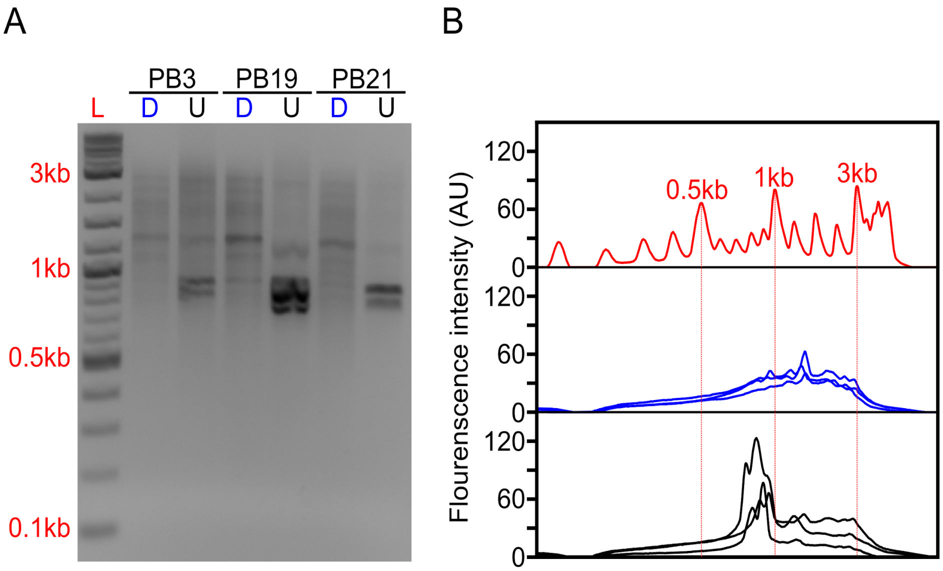


Figure 3.2: Long-DASH depletes hemoglobin from full-length cDNA.
A) Depleted (D) or undepleted (U) cDNA was visualized on a 2% Agarose gel. DNA ladder (L) suggests highly abundant cDNA species - putatively hemoglobin around ~700bp. B) Fluorescence analysis of the gel by ImageJ (Rasband 2011) further emphasizes the difference between depleted (blue) and undepleted (black) cDNA pool. Select size markers in the DNA ladder (red) are indicated.

Long-DASH is compatible with Smart-seq2 library preparation and does not distort cDNA composition

Next, we aimed to validate whether Long-DASH truly depletes hemoglobin transcripts in the cDNA pool and can be used for Illumina's short-read RNA-seq sequencing platform. To show this we prepared independent Tn5 based Smart-seq2 sequencing libraries for each depleted and undepleted cDNA pool. We then sequenced the Smart-seq2 libraries on a multiplexed Illumina HiSeq X 2 x151 bp run. We generated ~20 million reads for depleted and ~60 million reads for undepleted samples. By sequencing the undepleted samples deeper, we reasoned that the non-hemoglobin genes should receive equivalent read coverage in depleted and undepleted samples. This allowed us to perform a side by side comparison of the depleted and non-depleted samples to ensure no off-target effects.

First, we analyzed the resulting sequencing data using a custom kmer based approach to estimate the number of reads originating from hemoglobin transcripts. In the undepleted cDNA pools 48-68% of reads were scored as originating from hemoglobin transcripts. In depleted samples this was reduced to 1-4% reads (Fig. 3.3A). As a consequence, at the same read depth, RPM values for non-hemoglobin genes increased by a factor of 3 on average.

Second, to show that the depletion of hemoglobin transcripts did not distort the rest of the cDNA pool, we aligned the reads to the polar bear genome and quantified the expression of all previously annotated genes. We observed that when reads aligning to the hemoglobin loci were included in the analysis, the reads aligning to the few hemoglobin loci in our undepleted samples skewed the RPM calculations. By sequencing undepleted samples to great depth, this allowed us to exclude hemoglobin from quantification of gene expression while matching non-hemoglobin read depth of depleted samples. This analysis showed that the overall gene expression patterns were not dramatically distorted between depleted and undepleted samples. The three polar bear samples showed a Pearson r-value of 0.97-0.98 (Fig. 3.3B) when the gene expression values of depleted and undepleted samples were compared and reads aligning to hemoglobin loci were discarded.

Next we checked for genes whose expression was systematically affected by depletion. No genes were downregulated more than 4-fold in all three polar bear samples suggesting that there was no strong systematic off-target effects using the Cas9 based depletion. We did however find 151 genes out of ~12000 expressed genes to be upregulated by at least 4-fold in all 3 polar bears suggesting that Cas9 based depletion and subsequent second PCR amplification have had a systematic impact on a number of genes. We then investigated whether this effect would affect differential expression analysis between depleted samples. To this end, we calculated gene expression differences for each pair of polar bears twice, once pre- and once post-depletion. We then compare the pre- and post-depletion gene expression differences and found that, while depletion does introduce differences in the upregulated genes, these effects appear to be small, random in direction, and similar to a random selection of genes with similar expression levels (Supplementary Fig. S3.2).

Overall, this indicates that the depletion of hemoglobin from full-length cDNA pools was successful, thereby freeing up the vast majority of sequencing reads to analyze the rest of the polar bear transcriptome. Although, the data suggests that a number of genes were systematically affected by depletion and additional PCR steps, further experiments including several technical replicates should enable differential expression analysis between depleted samples.

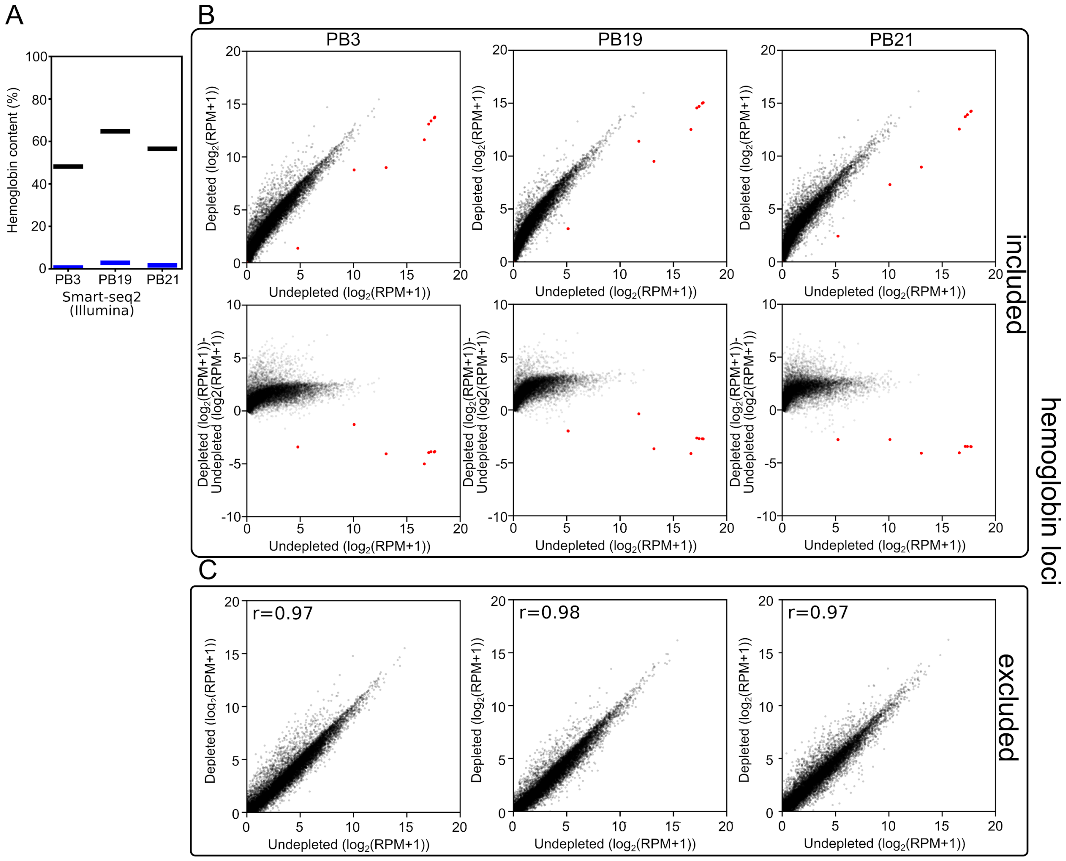


Figure 3.3: Long-DASH specifically targets hemoglobin from cDNA.
A) Hemoglobin content was measured in Smart-seq2 (Illumina) libraries of depleted (blue) or undepleted (black) cDNA pools. B(Top) and C) Scatterplots comparing gene expression in undepleted and depleted Smart-seq2 libraries of PB3, PB19, and PB21 with reads aligning to hemoglobin loci (red) (Liu et al. 2014) either included (B) or excluded (C). B(Bottom) Scatterplots showing log2(fold-change) between depleted and undepleted cDNA pools as calculated by [Depleted (log2(RPM+1))-Undepleted (log2(RPM+1))] with hemoglobin loci included in the RPM normalization.

Long-DASH is compatible with full-length cDNA sequencing methods

Having established the compatibility of Long-DASH with the short-read RNA-seq assay, we investigated whether we could generate a long-read data set from the depleted cDNA using our R2C2 approach. By incorporating R2C2 we can generate error-corrected full-length cDNA reads using long-read ONT sequencers. We used 5 partially multiplexed flowcells to generate ~3.8 million R2C2 consensus reads of 5 depleted cDNA pools – two Long-DASH replicates (R1 and R2) for PB3 and PB19 as well as a single Long-DASH run for PB21. The R2C2 reads we generated had a median accuracy of 94%, which is between 8-10% more accurate than standard ONT cDNA sequencing protocols (Supplementary Table S3.1).

We also generated ~5,000 R2C2 consensus reads of undepleted cDNA from one polar bear which allowed us to compare hemoglobin content and consensus read length distributions between depleted and undepleted samples (Fig 3.4). In the undepleted sample, the majority of R2C2 reads were of two distinct lengths, both around 700 bp, likely representing 79.3% of hemoglobin transcripts present in that sample. The 5 depleted samples showed a much more evenly distributed read length with a median hemoglobin content of 1.2% (0.6%-8.3%) (Fig. 3.4). Higher hemoglobin levels for R2C2 compared to Smart-seq2 based library preps (1-4%) using the same cDNA might be explained with R2C2 being somewhat biased towards transcripts between 500-1000 bp.

The median read length of the depleted samples was slightly below 1 kb which is in line with cDNA read length distributions published to date (Workman et al. 2018). This means that despite the less than ideal conditions for RNA integrity given difficult field conditions and the lag time between sample collection and processing, the analyzed RNA molecules were largely intact.

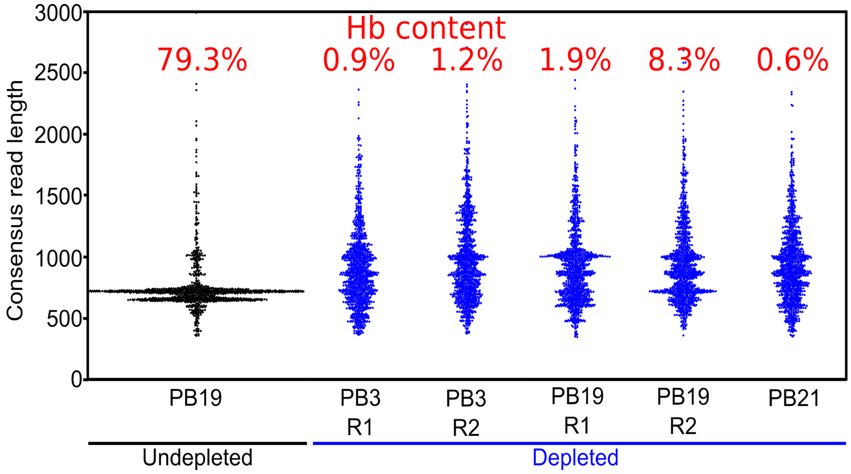


Figure 3.4: Long-DASH depletes hemoglobin from cDNA.
Length distribution of R2C2 consensus reads is shown as swarmplots in the indicated samples. Independent Long-DASH replicates (R1 and R2) were performed for samples PB3 and PB19 but not PB21. Percent of Hemoglobin reads as determined with a kmer approach is given in red on top.

R2C2 reads of depleted full-length cDNA can refine transcriptome annotations

Next, we generated high confidence isoform-level information from our full-length cDNA to refine the currently available polar bear transcriptome annotation. To this end, we analyzed our 3.8 million R2C2 consensus reads using the Mandalorion pipeline we previously developed (Byrne et al. 2017). We aligned the R2C2 reads to the polar bear genome sequence (Liu et al. 2014) using minimap2. These alignments, together with previously known individual splice sites (Genomic Resources Development Consortium et al. 2014; Liu et al. 2014), then serve as input into our Mandalorion pipeline which processes read alignments into high-confidence isoforms.

The Mandalorion pipeline first complements known splice sites with new splice sites it identifies de novo from R2C2 read alignments. It then groups R2C2 reads based on the splice sites they use. Pairs of transcription start sites (TSSs) and polyA sites are then determined for each group to identify full-length isoforms. Two additional processing steps were performed whereby isoforms were excluded if they were fully contained within longer isoforms or unspliced. This was to ensure removal of any non-full-length isoforms that may result from RNA degradation, as well as isoforms potentially caused by DNA contamination, respectively. In total, this analysis produced 5,831 high-confidence isoforms with a median accuracy of 99.1%.

We then classified these 5831 high-confidence spliced isoforms using the Sqanti algorithm (Tardaguila et al. 2018) that determines what relationship an experimentally determined isoform has to genes and isoforms in a reference annotation (Fig. 3.5). As a reference, we downloaded 28,880 known and predicted mRNA sequences from NCBI by selecting “RefSeq” and “mRNA” filters in the NCBI Nucleotide database, most of which are based on the NCBI Ursus maritimus Annotation Release 100 catalog of polar bear mRNA sequences (Pruitt et al. 2014).

1239 of the 5831 Mandalorion isoforms were classified as “novel_not_in_catalog” (NNC) which means that they overlapped a known gene but contained at least one unannotated splice site. In-depth analysis of this NNC group found that they contained a total of 521 new exons. In addition to R2C2 read coverage, Smart-seq2 read coverage was elevated in these new exons providing additional evidence for their inclusion in transcripts. Further, 1301 isoforms were classified as "novel_in_catalog" (NIC), which means that they overlapped a known gene and used only annotated splice sites but at least once as part of a previously unannotated splice junction. In total we observed 2540 (1239 NNC and 1301 NIC) new isoforms with unannotated exon configurations. An additional 1893 isoforms were classified and "full_splice_match" (FSM) which means that their splice junctions matched an annotated isoform completely but it doesn't mean that TSS and polyA sites also matched this isoform. In-depth analysis of the putative full-length NNC, NIC, and FSM isoform groups identified 2885 new TSSs and 1817 new polyA sites. R2C2 read coverage declined rapidly at TSSs and polyA sites providing clear evidence for their validity. Smart-seq2 read coverage was elevated inside TSS and polyA sites but declined slowly towards the respective features which is characteristic for standard short-read Illumina data (Fig.3.5). This is not surprising as short-read based protocols have to be specifically designed to capture those features (Salimullah et al. 2011; Cole et al. 2018; Ruan and Ruan 2011). So, while this data validates the existence of these features, it cannot be used for confirming their exact location.

Finally, 769 isoforms were classified as "incomplete_splice_match" (ISM) which means that they contain a subset of splice junctions of an annotated isoform. While these isoforms could represent real, shorter transcripts, they might also represent experimental artifacts so we excluded them from TSS and polyA analysis.

Considering RefSeq mRNA sets are in part based on deep short-read data and computational annotation, we did not expect to discover many entirely new gene loci. However, 509 of the 5831 isoforms were classified as "intergenic" (IG) which means that they did not overlap with any annotated gene locus. By determining which of these isoforms overlapped with each other, we identified 176 new gene loci.

Overall, this analysis dramatically refined our isoform-level knowledge of the whole blood polar bear transcriptome (Fig. 3.5). To make this knowledge straightforward to use for future analysis, we have generated a gtf annotation file containing RefSeq mRNA entries merged with our R2C2/Mandalorion isoforms.

How these new isoforms and isoform features have improved the current annotation can be seen clearly in these three following examples. In the RBX1 gene, we discovered 10 new isoforms containing new TSSs and polyA sites, several of which were associated with new terminal first or last exons (Fig. 3.6A). In the GMFG gene, we similarly identified new isoforms containing unannotated internal and terminal exons, intron retention events, TSSs, and polyA sites (Fig. 3.6B). Finally, we discovered a new gene locus that contains two isoforms and is entirely absent in the polar bear RefSeq mRNA set. However, aligning the two isoforms to the Panda genome (R. Li et al. 2010) resulted in unique matches to the CCDC72 gene (Fig. 3.6C).

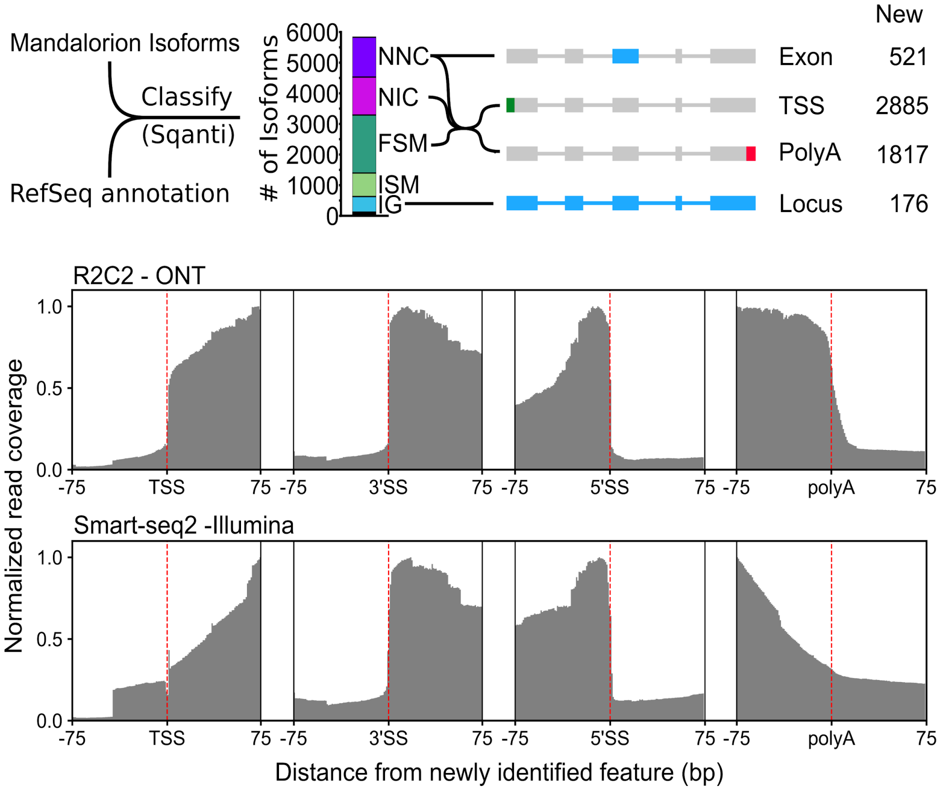


Figure 3.5: R2C2/Mandalorion identifies new isoform features in the polar bear transcriptome.

(Top) General workflow for comparing RefSeq mRNAs and Mandalorion isoforms is shown on the left. RefSeq mRNAs were aligned to the polar genomes using minimap2 and converted to gtf format to create a reference annotation. Isoforms determined by Mandalorion were then classified using this reference annotation using the sqanti_qc algorithm. Isoforms were classified as Novel_not_in_catalog (NNC), Novel_in_catalog (NIC), Full_splice_match (FSM), Incomplete_splice_match (ISM) and Intergenic (IG). New transcriptome features were then determined based on the minimap2 alignments of isoforms in the indicated categories. (Bottom) R2C2 and Smart-seq2 read coverage around newly identified TSS, the splice sites (3' and 5') of newly identified exons, and newly identified polyA sites.

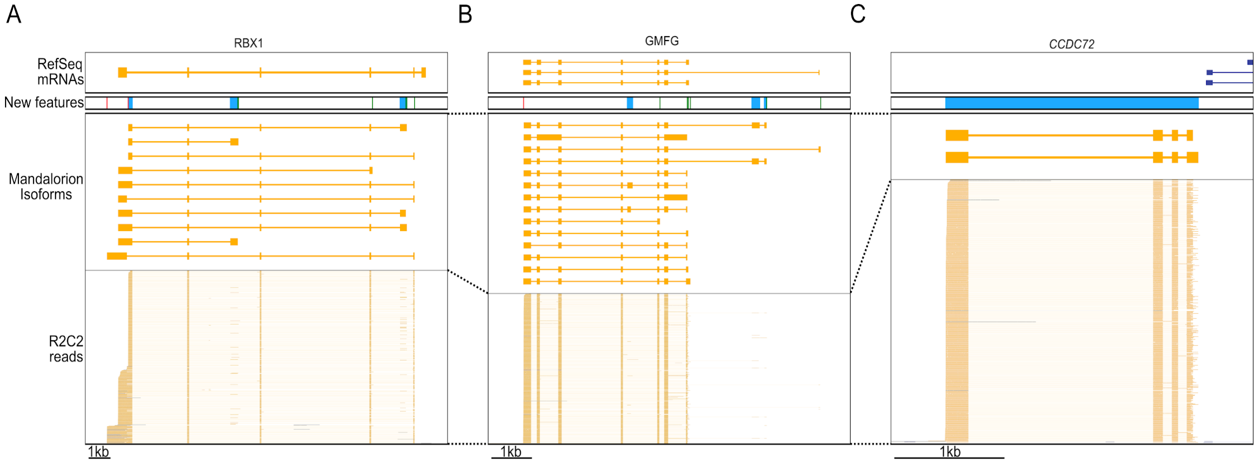


Figure 3.6: R2C2/Mandalorion refine transcriptome annotations.

Genome Browser views of the RBX1 locus (A), the GMFG locus (B), a locus likely corresponding to the CCDC72 gene not yet included in the RefSeq mRNA set (C). From top to bottom, 1) RefSeq mRNAs alignments, 2) new features based on Mandalorion isoforms (green: TSS, red: polyA site, blue: new exon or locus), 3) Mandalotion isoforms, and 4) R2C2 reads. Plus, strand alignment are in blue, minus strand alignments in orange.

Discussion

To better understand how humans and environmental perturbations impact threatened or endangered species, it is critical to understand changes in transcriptome dynamics. Fluctuations at the molecular and cellular level are sensitive indicators of environmental change (Brown et al. 2017; Kim et al. 2011); they are analogous to veterinary medicine where blood transcriptomes serve as proxies for identifying health status, disease, and exposures to environmental toxicants (Lv et al. 2018; McLoughlin et al. 2014; Burgess et al. 2012; Watson et al. 2017). Changes at the transcriptome level may also be useful indicators of ecological specialization, and therefore useful to design strategies for species management and conservation (Supple and Shapiro 2018). However, existing approaches to generate transcriptome data from whole blood RNA are either specifically designed for short-read sequencing (DASH) or human samples (commercial hemoglobin depletion kit like GLOBINclear) and therefore lack a cost-effective approach for analyzing isoform-level transcriptomes of non-model organisms.

Any study investigating whole blood transcriptomes using short- or long-read sequencing will greatly benefit from the Long-DASH method. Long-DASH effectively and economically depletes hemoglobin transcripts from whole blood full-length cDNA which can then be sequenced with short- or long-read sequencing. We validated Long-DASH by depleting hemoglobin transcripts from polar bear whole blood cDNA pools and generated deep short-read Smart-seq2 RNA-seq data as well as 3.8 million R2C2 full-length cDNA consensus reads. We processed the 3.8 million full-length R2C2 reads to identify close to 6000 high confidence isoforms which we then used to refine and improve the polar bear whole blood transcriptome annotation.

In addition to polar bear hemoglobin transcripts, the sgRNAs designed for this study will also target human hemoglobin transcripts making them useful for basic research as well as clinical applications in cancer biology and disease diagnosis (Fig. S1) (Valk et al. 2004; Borovecki et al. 2005; Gervasoni et al. 2008; Morey et al. 2016). Further, the sgRNA sequences used in Long-DASH can be easily adapted to any organisms or gene. The ease and adaptability places Long-DASH at an advantage over "as-is" commercial kits like GLOBINclear (Ambion), which promises >95% of depletion of human and mouse hemoglobin transcripts, but fails to efficiently deplete hemoglobin transcripts from pig whole blood RNA samples (Choi et al. 2014).

Since cDNA can be generated from femtogram levels of polyA-RNA, Long-DASH requires very little RNA input compared to RNA pull-down methods. This allows the investigator to gather small samples, or only process small aliquots of existing samples, thereby maximizing the usefulness of each sample collection and minimizing harm to animals. In its current state depletion by Long-DASH is still somewhat variable, resulting in hemoglobin levels from 0.6%-8.3%. While still a large improvement compared to the undepleted samples, future work on the method will center on removing this variability through either longer incubation times or higher number or concentration of sgRNA probes and the Cas9 protein. It may also be beneficial to measure depletion success by qPCR before sequencing a depleted cDNA pool.

Going forward, the Long-DASH depletion method and the R2C2 long-read sequencing method will form a very powerful combination for transcriptome analysis and annotation from whole blood samples and beyond. The transcriptomes of many tissues contain several highly abundant transcripts that represent >50% of all transcript molecules (Mure et al. 2018). A set of sgRNAs targeting any abundant transcripts can be easily generated, making Long-DASH conducive for surveying other tissues as well. Specifically, depleted full-length cDNA libraries can be sequenced using our R2C2 method, which currently represents the most powerful combination of throughput and accuracy in the long-read sequencing field. Our most recent R2C2 run emphasizes this by generating ~1,000,000 R2C2 reads at a median accuracy of 97.5% on a single ONT MinION flowcell at a cost of ~$650 (Table S1). This represents an increase in accuracy of >10% over standard ONT cDNA sequencing and 10-times more complete reads than the PacBio Sequel at the same cost. Combining our Long-DASH and R2C2 methods therefore brings the exhaustive annotation of non-model organisms within reach.

Materials and Methods

Sample Collection/RNA Extraction from Whole Blood

Permits for field operations and animal care were provided by the Government of Greenland (Permit numbers 2015-110281 and 2017-5446). Polar bear whole blood samples were collected in PAXgene Blood RNA tubes (PreAnalytiX GmbH, BD Biosciences, Mississauga, ON, Canada). Total RNA was isolated from whole blood (2.5mL) thawed at room temperature for 2 hours prior to using the PAXGene RNA extraction kit (Qiagen, Chatsworth, CA, USA) according to manufacturer's protocol. All samples were DNAse (Qiagen) treated and eluted in 50µL. The RNA yield and purity were accessed using a NanoDrop 8000 UV Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). RNA quantities ranged from 110 - 310 ng/µL and the A260/280 ratio values were ≥ 2.0. Human whole blood RNA was purchased from Zyagen Labs (NC1453913).

Full-length cDNA Generation

RNA was reverse transcribed (RT) using Smartscribe Reverse Transcriptase (Clonetech). We generated full-length cDNA using a modified Smart-seq2 approach (Cole et al. 2018). During the RT reaction a template-switch oligo and an oligodT primer was used to select for polyA+ RNA (Table S2). The RT reaction was performed in 10 µL reactions with an input of 70 ng of RNA and took place at 42°C for 1 hour. After cDNA synthesis, 1 µL of 1:10 dilutions of RNAse A (Thermofisher) and Lambda Exonuclease (NEB) were added and incubated at 37°C for 30 minutes. Following the incubation, an amplification step was performed in 25 µL final volumes using KAPA Hifi ReadyMix 2X (KapaBiosystems) containing 1 µL of the ISPCR primer (10 µM) primer. Samples were incubated at 95°C for 3 minutes, followed by 12 cycles of (98°C for 20 s, 67°C for 15 s, and 72°C for 4 minutes), with a final extension of 72°C for 5 minutes. Samples were purified using Agencourt AMPure XP SPRI beads (Beckman

Coulter) and eluted at 25 mL. The final cDNA product was then visualized on an agarose gel to confirm distribution (Fig. 3.2).

In-vitro Preparation of CRISPR/Cas9

SpCas9-2xNLS was purified based on the protocol described in (Jinek et al. 2012). Briefly, a plasmid encoding His_6MBP-SpCas9-2xNLS (Addgene plasmid #69090) was transformed into Rosetta2(DE3) *E. coli* cells. Cultures were grown at 37ºC in 2YT medium with shaking until they reached an OD_{600} of ~0.6, and then placed on ice for 5 minutes before adding IPTG to a final concentration of 0.25 mM; cultures were then grown overnight at 18ºC with shaking. Cell pellets were harvested by centrifugation, and then lysed in an Avestin cell extruder in Ni-A buffer (20 mM Tris pH 8.0, 500 mM NaCl, 5% vol/vol glycerol, 25 mM imidazole) with EDTA-free protease inhibitors (Pierce). Clarified supernatants were purified by gravity column on Ni-NTA agarose (QIAGEN) using Ni-A buffer to load and wash, and Ni-B buffer (20 mM Tris pH 8.0, 500 mM NaCl, 5% vol/vol glycerol, 250 mM imidazole) to elute. Peak fractions were concentrated in an Amicon Ultra spin concentrator with a 30 kDa molecular weight cutoff at 4ºC, and then loaded onto a 50 mL HiPrep Desalting Column (GE Healthcare) pre-equilibrated with 17% IEX-B (IEX-A buffer: 20 mM HEPES pH 7.5, 150 mM KCl, 5% vol/vol glycerol; IEX-B 20 mM HEPES pH 7.5, 1 M KCl, 5% vol/vol glycerol). The flow-through was then loaded onto a 2 mL HiTrap SP column (GE Healthcare) in 17% IEX-B buffer. After thoroughly washing the column in 17% IEX-B, the protein was eluted with a linear gradient from 17-50% IEX-B. Peak fractions were pooled and loaded onto a Superdex 200 16/60 column (GE Healthcare) pre-equilibrated in 20 mM HEPES pH 7.5, 150 mM KCl, 1 mM DTT, 10% vol/vol glycerol. Peak fractions were concentrated in an Amicon Ultra spin concentrator with a 30 kDa molecular weight cutoff at 4ºC until a concentration of 40 μM, which was estimated using the calculated molar extinction coefficient of 120,575 M^{-1} cm^{-1}. The protein was aliquoted into small volumes (10 μL), quick frozen in liquid nitrogen, and stored at -80ºC.

sgRNA Design and Construction

Other studies have shown that sgRNAs designed between 17-20 bp showed increased efficacy (Ren et al. 2014; Fu et al. 2014). As a result, the sgRNAs were designed between 17-20 bp in length. sgRNAs were designed to target hemoglobin transcripts in human and polar bear. A multi-sequence alignment was performed on the human and polar bear annotated HBA and HBB gene transcripts to find conserved regions using the Clustal Omega tool (W. Li et al. 2015; Sievers et al. 2011; McWilliam et al. 2013) (Fig. S3.3). Regions with high homology were chosen for sgRNA design. sgRNAs that did not share complete homology were designed to contain degenerate bases to ensure compatibility across species using the same sgRNA (Fig. S3.4). sgRNA specificity was determined by using BLAST (Altschul et al. 1990). One sgRNA was designed even though the N-GG (PAM motif) had been lost in the human but was still kept in the pool for the polar bear depletion (Fig. S3.3). A total of 16 sgRNAs were designed to target alpha and beta hemoglobin transcripts. The target oligos were then constructed into sgRNAs as previously described (Ren et al. 2014). Single stranded oligos were designed to contain a T7 promoter attached to each sgRNA sequence (IDT) followed by the first 22 bases of the tracrRNA sequence (Fig. S3.3). The complementary tracrRNA and single stranded oligo were annealed and extended to form a dsDNA product containing the T7-sgRNA and tracrRNA template. The template was then used for in-vitro transcription using the HiScribe T7 High Yield RNA synthesis kit (NEB). The in-vitro transcription reaction was carried out at 37°C for 16 hrs. The in-vitro transcribed RNA was then purified using MEGAclear Transcription Clean-Up Kit (Invitrogen). The final sgRNA product was then checked for purity and quantified using NanoDrop 8000 UV Spectrophotometer (Thermofisher). All sgRNAs were then pooled at equal molar concentrations and stored in single-use aliquots at -80°C.

CRISPR/Cas9 Treatment

Since it has been predicted that human whole blood samples can contain up to ~50-80% of hemoglobin transcripts of the total sample (Field et al. 2007; Mastrokolias et al. 2012), we calculated the ratio of sgRNA and Cas9 molar amounts to sample based upon this assumption. According to the DASH protocol it was determined that 150-fold of Cas9 and 1500-fold of sgRNA should be sufficient (Gu et al. 2016). All cDNA samples were quantified by Qubit using the dsDNA HS assay kit (Thermofisher) to calculate the molar amounts. To calculate the expected molar amounts we use the following formula:

$$nM = [DNA\ (ng/\mu L)] \div (660g/mol\ x\ size\ of\ hemoglobin\ transcripts\ in\ bp)$$

Once the molar amounts were determined, the ribonucleoprotein (RNP) complex was formed by adding the 150-fold Cas9 and 1500-fold sgRNA excess amount with 1.0 µL of the 10X Cas9 Buffer (final concentration 50 mM Tris pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$, and 1 mM TCEP) and incubated for 25°C for 10 minutes. Following the 25°C incubation, the calculated cDNA amount was then added (final volume of 10 µL) and incubated at 37°C for 4 hrs to overnight. After the Cas9 depletion, 1 µL of Proteinase K and RNAse A were added to inactivate the Cas9 and remove excess sgRNAs from the reaction and incubated at 37°C for 15 minutes and 95°C for 15 minutes. It is critical that the Proteinase K is deactivated properly as the samples are immediately used for amplification. Treated samples were PCR amplified (95°C for 3 minutes, followed by 13 cycles of (98°C for 20 s, 67°C for 15 s, and 72°C for 4 minutes) followed by a final extension of 72°C for 5 minutes). PCR was performed using KAPA Hifi ReadyMix 2X (KapaBiosystems) and 1 µL of the (10 µM) ISPCR primer. The amplified product was then purified using SPRI beads to remove everything below 500 bp. Selecting against cDNA below 500 bp ensured that all cut hemoglobin products were removed before making the Tn5 libraries. The depleted cDNA product was visualized on a 1-2% agarose gel to confirm depletion. Once confirmed, the depleted cDNA product was then prepped for either Illumina or Nanopore sequencing.

R2C2 Library Preparation and ONT sequencing

To prepare R2C2 libraries ~30 ng of the depleted cDNA was used. The R2C2 libraries were made as previously described (Volden et al. 2018). Briefly, an equal concentration of splint to cDNA were combined (30ng of depleted cDNA and 30ng of our (~200 bp) DNA splint). The full-length cDNA was then circularized using the 2X NEBuilder Hifi DNA Assembly Mix (NEB). The reaction took place at 50°C for 1 hour per manufacturer protocol. Once the full-length cDNA was circularized, linear ssDNA and dsDNA was digested by adding 3μL each of Lambda Exonuclease, Exonuclease I and Exonuclease III (all NEB) and incubated at 37°C overnight. We performed the longer incubation overnight to ensure complete digestion. After digestion, the sample was further purified using SPRI beads and eluted in 30μL of ultrapure water. 30μL of sample was then split into three reactions containing 10μL each for the Phi29 amplification. The Phi29 amplification took place in a reaction volume of 50μL containing 5μL of 10X Buffer, 2.5μL of 10uM each dNTPs, 2.5μL of random hexamers (10uM), 29μL of ultrapure water and 1μL of Phi29 Polymerase. The Phi29 reactions were incubated at 30°C for 16 hrs, 65°C for 15 minutes and held at 4°C. All three samples were pooled together and ultrapure water was added to make up the final volume to 300μL. The product was purified using SPRI beads with a 1:0.5 sample to bead ratio. This ratio was chosen as it removed all fragments < 2000 kb. The sample was then eluted in 90μL of ultrapure H20, 10μL of NEB2 Buffer (NEB) and 3μL of T7 endonuclease (NEB) and incubated at 37°C for 2 hrs to ensure complete debranching of the Phi29 product. The eluted sample was again purified using SPRI beads with a 1:0.5 sample to bead ratio. The product was eluted in 30μL and quantified using Qubit dsDNA HS kit (Thermofisher). The length distribution was verified on a 1% agarose gel prior to performing the ONT library prep.

For the library preparation ~1-2 μg of the final R2C2 product was converted into a ONT compatible library using the SQK-LSK109 kit according to ONT instructions with minor modifications. First, during the End Repair and A-tailing reaction we performed incubations for 30 min each at 20°C and 65°C instead of 5 min each. Second, we adjusted the ligation reaction time to 30 minutes at room temperature instead of 10 minutes per the protocol. We also found that loading between ~200-500 ng of the final library onto the flowcell was the most optimal. Loading more library resulted in severe loss in throughput as can be seen for the R2C2 runs PB3_depleted_R1 and PB19_depleted_R1 (Table S3.2). R2C2 libraries were sequenced on a MinION device using the 48hr sequencing protocol using the FLO-Min106 R9.4 Rev D chemistry flowcells. All reads were basecalled with Albacore v2.1.3.

Smart-seq2 Library Preparation and Illumina Sequencing

Illumina libraries of the depleted and non-depleted samples were prepared using a tagmentation based method using our own Tn5 (Picelli, Faridani, et al. 2014). The Tn5 enzyme was custom loaded with Tn5ME-A/R and Tn5ME-B/R adapters (Table S3.2). The Tn5 reaction contained 5μL of the full-length cDNA product, 1μL of the loaded Tn5 enzyme, 10μL of ultrapure water and 4μL of the 5X TAPS-PEG buffer and incubated at 55°C for 7 minutes. After incubation, 5μL of 0.2% of Sodium Dodecyl Sulfate (SDS) was added to the product to inactivate the Tn5 enzyme. Due to the Tn5 generating gaps, 5μL of the Tn5 product had to be nick translated at 72°C for 5 mins.

The Tn5 product was then amplified using KAPA Hifi Polymerase (KAPA) with 10 cycles of PCR using (98°C for 10 s, 63°C for 30 s, 72°C for 2 min) with a final extension at 72°C for 5 min. The final reaction volume was 25µL and contained 0.5µL KAPA Hifi Polymerase (KAPA), 5µL of 5X Buffer, 0.8µL of dNTPs (10mM each), 11.7µL of ultrapure water, 5µL of the nick-translated product and 1µL each of Nextera_Primer_A and Nextera_Primer_B primers (Table S3.2). The amplified Tn5 libraries were then size selected from 300 - 800 bp on a 2% EX E-gel (Thermofisher) and purified using QIAquick gel extraction kit (Qiagen). The libraries were then pooled at equal concentrations and ran on a HiSeq X 2x151 bp run.

R2C2 read processing and isoform analysis

R2C2 consensus reads were generated from raw reads using the C3POa pipeline (https://github.com/rvolden/C3POa). C3POa identifies subreads in the raw reads and uses poaV2 (Lee, Grasso, and Sharlow 2002) and racon (Vaser et al. 2017) to determine a more accurate consensus of these subreads. The consensus reads were then aligned to the polar bear genome (Liu et al. 2014) using minimap2 (H. Li 2017) using standard setting and the *'-ax splice'* flag. The resulting sam files are converted to psl files using samtools (H. Li et al. 2009) and jvarkit samtopsl utility (Lindenbaum 2015).

The resulting psl, sam, and fasta files of all depleted samples were merged and used as input into the Mandalorion (https://github.com/rvolden/Mandalorion-Episode-II) pipeline to determine isoforms. To accomodate issues regarding RNA degradation and genomic DNA contamination, we integrated two new optional filter into

Mandalorion. We implemented the filtering of isoforms that are entirely contained within one other isoform, which indicates degraded input RNA molecules, and the filtering of unspliced isoforms which might stem from DNA contamination.

Accuracy of R2C2 reads and Mandalorion isoforms were determined using alignments in sam format containing md-strings and a custom script that calculates

Accuracy= Matches/(Matches+Mismatches+Indels)

Smart-seq2 read processing

Paired fastq files were downloaded from basespace and aligned to the polar bear genome using STAR with standard settings. The STAR index for the polar bear genome was built without a transcriptome reference because the gff file provided by (Liu et al. 2014) did not conform to gff standard (no “exon” features) and could therefore not be used. Read alignments in ordered bam format were converted to psl as described above.

Hemoglobin and gene expression quantification.

Hemoglobin content was determined through a kmer based counting method using a custom script. In short, all possible 10nt kmers were extracted from the sequence of hemoglobin alpha and beta transcripts. The presence of these kmers were then determined in each R2C2 or Smart-seq2 read from depleted and undepleted samples. Cutoffs for read assignments to hemoglobin were then determined by also analyzing R2C2 and Illumina reads of the GM12878 cell line which does not express hemoglobin.

Gene expression was determined using Smart-seq2 (Illumina) read alignments in psl format and a custom script. Reads aligning to hemoglobin loci were not counted towards total aligned reads in the RPM calculations.

Both script are available are available at https://github.com/christopher-vollmers/PB_scripts.

Data Visualization

Schematics were prepared using inkscape (https://inkscape.org). All others figures were prepared using python/matplotlib/numpy/scipy (Millman and Aivazis 2011; Jones, Oliphant, and Peterson 2001--; van der Walt, Colbert, and Varoquaux 2011; Hunter 2007)

Data Availability

All Illumina and ONT raw read data is available at SRA under Bioproject accession PRJNA514749.

Supplemental

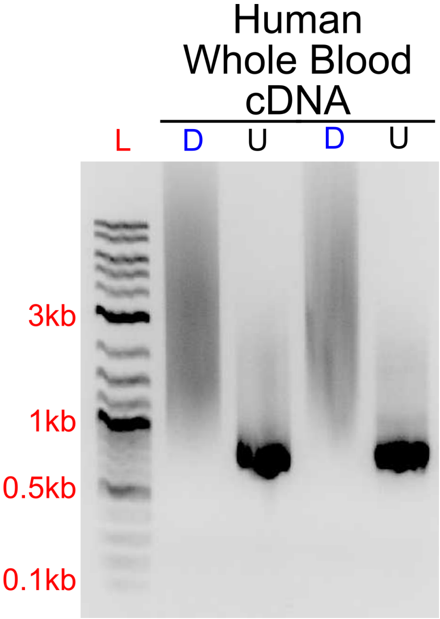


Figure S3.1: Long-DASH sgRNA also depletes human hemoglobin transcripts from full-lengthcDNA.

Technical replicates of depleted (D) or undepleted (U) human whole blood cDNA were visualized on an agarose gel. DNA ladder (L) suggests highly abundant cDNA species - putatively hemoglobin around ~700bp.

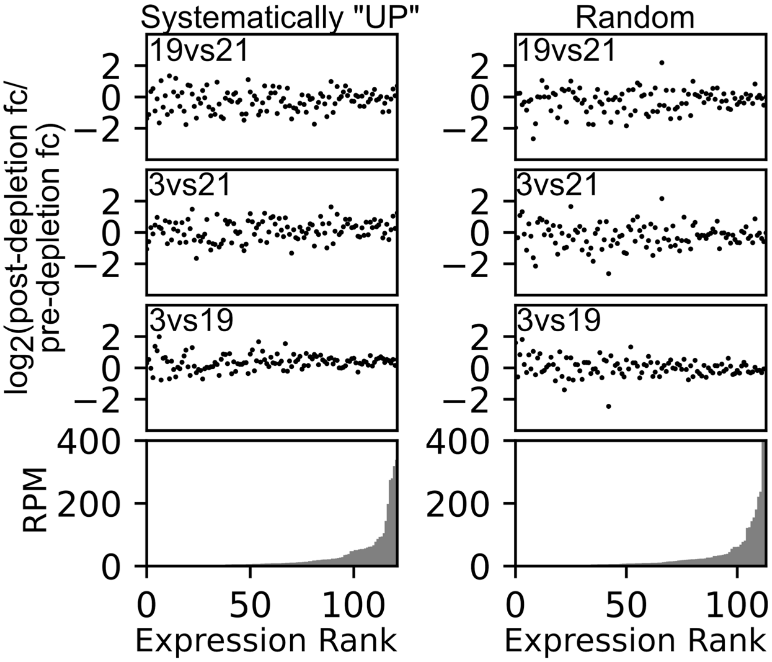


Figure S3.2: Depletion of hemoglobin affects expression levels within but not fold-change between samples

Expression levels(bottom) as well as the changes in differential expression between polar bears pre- and post-depletion are shown for genes that are upregulated in all 3 polar bears post-depletion (Systematically “UP”) and a random selection of genes with similar expression distribution (Random).

A) Hemoglobin, alpha (HBA)

XM_008696690.1 --------------------CCGCCCCGCACATTTCTGGTCCTCACAGACTCAGAAAGAA 40
NM_000558.4
CATAAACCCTGGCGCGCTCGCGGCCCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAA
60
NM_000517.4
CATAAACCCTGGCGCGCTCGCGGGCCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAA
60
* * ** **** ********* ************ ****

XM_008696690.1
GCCACCATGGTGCTGTCTCCCGCCGACAAGAGCAACGTCAAGGCCACCTGGGATAAGATC
100
NM_000558.4
CCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTC
120
NM_000517.4
CCCACCATGGTGCTGTCTCCT**GCCGACAAGACCAACGTCA**AGGCCGCCTGGGGTAAGGTC
120
****************** ********** ************* ****** **** **

XM_008696690.1
GGCAGCCACGCTGGCGAGTATGGCGGCGAGGCTCTGGAGAGGACCTTCGCGTCCTTCCCC
160
NM_000558.4
GGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCC
180
NM_000517.4
GGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCC
180
*** ***************** * ***** ********** *** **********

XM_008696690.1
ACCACCAAGACCTACTTCCCCCACTTCGACCTGAGCCCTGGCTCCGCCCAGGTCAAGGCC
220
NM_000558.4
ACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGC
240
NM_000517.4
A**CCACCAAGACCTACTTCCC**GCACTTCGACCTGAGCCACGGCTCTGCCCA**GGTTAAGGG**
C 240
******************* **************** ***** ******** **** *

XM_008696690.1
CACGGCAAGAAGGTGGCCGACGCCCTGACCACCGCCGCAGGCCACCTGGACGACCTGCCG
280
NM_000558.4
CACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCC
300
NM_000517.4
CACGG**CAAGA**AGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCC

```
300
                    ************************ ******* *****  *  *** ******** ****

XM_008696690.1
          GGCGCCCTGTCCGCTCTGAGCGACCTGCACGCGCACAAGCTGCGAGTGGACCCGGTCAAC
340
NM_000558.4
          AACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAAC
360
NM_000517.4
          AACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAA
C    360
                     *** ******** ************************** ** ***************

XM_008696690.1
          TTCAAGTTCCTGAGCCACTGCCTGCTGGTGACCCTGGCCAGCCACCACCCCGCGGAGTTC
400
NM_000558.4
          TTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
420
NM_000517.4
          TTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
420
                    ****** **** ***************************  ***** ****** ******

XM_008696690.1
          ACCCCTGCCGTCCACGCCTCCCTGGACAAGTTCTTCAGCGCCGTGAGCACCGTGCTCACC
460
NM_000558.4
          ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACC
480
NM_000517.4
          ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACC
480
                    ******** ** ********************* *  * ********************

XM_008696690.1
          TCCAAATACCGTTAAGCTGGAGCCGCGCGACCCTCCCGCTCCCGGCCTGGGGCCTCTTGC
520
NM_000558.4
          TCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCC
540
NM_000517.4
          TCCAAATACCGTTAAGCTGGAGCCTCGGTAGCCGTTCCTCCTGCCCGCTGGGCCTCCCAA
540
                    ************************ **          ***  *     *   *******

XM_008696690.1   GC--------------TCCACGCGCCTGAACTTCCCGATCTTTGAATAAAGTCTGAGTGG
566
NM_000558.4
          CAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGG
600
NM_000517.4      CGGGCCCTCCTCCCCTCCTTGCACC-
```

```
GGCCCTTCCTGGTCTTTGAATAAAGTCTGAGTGG    599
                                   *    *    *  *  * **********************

XM_008696690.1  GCTGCAG--------------------    573
NM_000558.4     GCGGCAAAAAAAAAAAAAAAAAAAAAAA    627
NM_000517.4     GCAGCAAAAAAAAAAAAAAAAAAA----    622
                ** **
```

B) Hemoglobin, beta (HBB)

```
XM_008709611.1
          GAGCAGGGCCAGCTGCTGCTTATACTTGCTTCTGACACAACCGTGTTCACTAGCAACCAC
60
NM_000518.4     ---------------------ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTC    39
                                     * * **************** **************** *

XM_008709611.1
          AAAGAGACACCATGGTGCATCTGACTGGTGAGGAGAAGTCTCTCGTCACCGGCCTGTGGG
120
NM_000518.4
          AAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
99
                *** **********************  *************  *** ** * ********

XM_008709611.1
          GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTTCTGGTTGTCTACC
180
NM_000518.4
          GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTAC
C    159
                ********************************************** ***** *******

XM_008709611.1
          CCTGGACTCAGAGGTTCTTTGACTCCTTTGGGGACCTGTCCTCTGCTGATGCTATTATGA
240
NM_000518.4
          CTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGG
219
                * ***** ************** *********** ****** ** ******** *****

XM_008709611.1
          ACAACCCCAAGGTCAAGGCCCATGGCAAGAAGGTGCTGAACTCCTTTAGTGATGGCCTGA
300
NM_000518.4
          GCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGG
279
                 ****** ***** ***** *********** *****     *****************

XM_008709611.1
          AGAATCTGGACAACCTCAAGGGCACCTTTGCTAAGCTGAGCGAGCTGCACTGTGACAAGC
```

360
NM_000518.4
CTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGC**ACTGTGACAAGC**
339
* ************************** * ***** ******************

XM_008709611.1
TGCACGTGGATCCCGAGAACTTCAAGCTCCTGGGCAACGTGCTGGTGTGTGTGCTGGCTC
420
NM_000518.4
TGCACGTGGATCCTGAGAACTTCAG**GCTCCTGGGCAACGTGC**TGGTCTGTGTGCTGGCCC
399
************* ********** ******************** *********** *

XM_008709611.1
ACCACTTTGGCAAAGAGTTCACCCCTCAGGTGCAGGCTGCCTATCAGAAGGTGGTGGCTG
480
NM_000518.4
ATCACTTTGGCAAAGAATTCACCCCACCAGTG**CAGGCTGCCTATCAGAAAG**TGG**TGGCT**
G 459
* ************** ******** * ******************** **********

XM_008709611.1
GTGTGGCCAACGCCCTGGCCCACAAGTACCACTGAGCTCCTGGCCTGTTTCCTGGTGATC
540
NM_000518.4
GTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTC
519
******* ** ***************** **** *****   **   *   **

XM_008709611.1 CCTG-
GAAGACCCTGTTCCCCTAAATTCTATCTTCTGAACTGGGGGAAATAATGTCCACC 599
NM_000518.4
TATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGC
579
*   * * *** *   **   ** ** * ** ** ********** ** ***   *

XM_008709611.1 ATCAAGGGTATGGTTTCTGCCTAATAAAGAACCTTCAGCTCAA---- 642
NM_000518.4 CTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC 626
* ** * *** ************** *** ** * * *

Figure S3.3. Alignment of orthologous HBA and HBB mRNA sequences in human and polar bear. Multi-sequence alignment from Clustal Omega v1.2.4. * indicates a match. Underline and bold indicates target sequences used for sgRNA design for Globin depletion. Red indicates (N-GG) PAM sequence a) Hemoglobin, alpha (HBA) b) Hemoglobin, beta (HBB).

A) Construction of sgRNAs

#1 tracrRNA oligo (in Reverse Orientation/Anti-sense)
<-------------------tracrRNA-------------------------------- <-------Primer----->
5'–
AAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTT
GCTATTTCTAGCTCTAAAAC–3'

The oligo above is a universal tracrRNA template which allows you to generate full sgRNA templates with any target sequence oligo as long as the target sequence oligo meets the following requirements below:
1. The oligo contains the reverse complement of the primer sequence on the 3' end.
2. the oligo contains 'GG' on the 3' end of the target sequence for T7 RT.

#1 HBA1/HBA2 target oligos based upon mRNA prediction of Polar Bear/Human genes.

<--------T7----------------><--target_sequence--><-------Primer------>
oligo 1 5'-GAAATTAATACGACTCACTATAGG AAGGSCCACGGCAAGAAGG
GTTTTAGAGCTAGAAATAGC-3'
<--------T7----------------><--target_sequence--><-------Primer------>
oligo 2 5'-GAAATTAATACGACTCACTATAGG CACTGCCTGCTGGTGACCC
GTTTTAGAGCTAGAAATAGC-3'
<--------T7----------------><--target_sequence--><-------Primer------>
oligo 3 5'-GAAATTAATACGACTCACTATAGG GGTYAAGGSCCACGGCAAGA
GTTTTAGAGCTAGAAATAGC-3'
<--------T7----------------><--target_sequence--><-------Primer------>
oligo 4 5'-GAAATTAATACGACTCACTATAGG ACCTCCAAATACCGTTAAGC
GTTTTAGAGCTAGAAATAGC-3'
<-------T7---------------><--target_sequence--><-------Primer------>
oligo 5 5'-GAAATTAATACGACTCACTATAGG GCCGACAAGASCAACGTCA
GTTTTAGAGCTAGAAATAGC-3'
<-------T7---------------><--target_sequence--><-------Primer------>
oligo 6 5'-GAAATTAATACGACTCACTATAGG GGGAAGTAGGTCTTGGTGGTG
GTTTTAGAGCTAGAAATAGC-3'
<-------T7---------------><--target_sequence--><-------Primer------>
oligo 7 5'-GAAATTAATACGACTCACTATAGG TCCTRAGCCACTGCCTGC
GTTTTAGAGCTAGAAATAGC-3'
<-------T7---------------><--target_sequence--><-------Primer------>
oligo 8 5'-GAAATTAATACGACTCACTATAGG CAGGTCGCTCAGRGCGGACA
GTTTTAGAGCTAGAAATAGC-3'

#2 HBB target oligos based upon mRNA prediction of Polar Bear/Human gene.

<-------T7----------------><--target_sequence--><-------Primer------>
oligo 1 5'-GAAATTAATACGACTCACTATAGG CACTGTGACAAGCTGCACG
GTTTTAGAGCTAGAAATAGC-3'

<-------T7----------------><--target_sequence--><-------Primer------>
oligo 2 5'-GAAATTAATACGACTCACTATAGG GAAGTTGGTGGTGAGGCCCT
GTTTTAGAGCTAGAAATAGC-3'
<-------T7----------------><--target_sequence--><-------Primer------>
oligo 3 5'-GAAATTAATACGACTCACTATAGG CAGGCTGCCTATCAGAARG
GTTTTAGAGCTAGAAATAGC-3'
<-------T7----------------><--target_sequence--><-------Primer------>
oligo 4 5'-GAAATTAATACGACTCACTATAGG GCAACCWCAAASAGACACCA
GTTTTAGAGCTAGAAATAGC-3'
<-------T7----------------><--target_sequence--><-------Primer------>
oligo 5 5'-GAAATTAATACGACTCACTATAGG GAGGTTCTTTGABTCCTTTG
GTTTTAGAGCTAGAAATAGC-3'
<-------T7----------------><--target_sequence--><-------Primer------>
oligo 6 5’-GAAATTAATACGACTCACTATAGG AAGGTGAACGTGGATGAAGT
GTTTTAGAGCTAGAAATAGC-3'
<-------T7----------------><--target_sequence--><-------Primer------>
oligo 7 5’-GAAATTAATACGACTCACTATAGG GCTCCTGGGCAACGTGC
GTTTTAGAGCTAGAAATAGC-3'
<-------T7----------------><--target_sequence--><-------Primer------>
oligo 8 5’-GAAATTAATACGACTCACTATAGG CAGAARGTGGTGGCTGGTG
GTTTTAGAGCTAGAAATAGC-3'

B) Target sequences and oligo sequences for hemoglobin depletion

HBA targets	HBB targets
(1) AAGGSCCACGGCAAGAAGG	(1) CACTGTGACAAGCTGCACG
Human- AAGGGCCACGGCAAGAAGG	Human- CACTGTGACAAGCTGCACG
Polar- AAGGCCCACGGCAAGAAGG	Polar- CACTGTGACAAGCTGCACG
(2) CACTGCCTGCTGGTGACCC	(2) GAAGTTGGTGGTGAGGCCCT
Human- CACTGCCTGCTGGTGACCC	Human- GAAGTTGGTGGTGAGGCCCT
Polar- CACTGCCTGCTGGTGACCC	Polar- GAAGTTGGTGGTGAGGCCCT
(3) GGTYAAGGSCCACGGCAAGA	(3) CAGGCTGCCTATCAGAARG
Human- GGTTAAGGGCCACGGCAAGA	Human- CAGGCTGCCTATCAGAAAG
Polar- GGTCAAGGCCCACGGCAAGA	Polar- CAGGCTGCCTATCAGAAGG
(4) ACCTCCAAATACCGTTAAGC	(4) GCAACCWCAAASAGACACCA
Human- ACCTCCAAATACCGTTAAGC	Human- GCAACCTCAAACAGACACCA
Polar- ACCTCCAAATACCGTTAAGC	Polar- GCAACCACAAAGAGACACCA
(5) GCCGACAAGASCAACGTCA	(5) GAGGTTCTTTGABTCCTTTG
Human- GCCGACAAGACCAACGTCA	Human- GAGGTTCTTTGAGTCCTTTG
Polar- GCCGACAAGAGCAACGTCA	Polar- GAGGTTCTTTGACTCCTTTG

(6) GGGAAGTAGGTCTTGGTGGTG (r) Human- GGGAAGTAGGTCTTGGTGGTG Polar- GGGAAGTAGGTCTTGGTGGTG	(6) AAGGTGAACGTGGATGAAGT Human- AAGGTGAACGTGGATGAAGT Polar- AAGGTGAACGTGGATGAAGT
(7) TCCTRAGCCACTGCCTGC Human- TCCTAAGCCACTGCCTGC Polar- TCCTGAGCCACTGCCTGC	(7) GCTCCTGGGCAACGTGC Human- GCTCCTGGGCAACGTGC Polar- GCTCCTGGGCAACGTGC
(8) CAGGTCGCTCAGRGCGGACA (r) Human- CAGGTCGCTCAGGGCGGACA (Lost PAM sequence) Polar- CAGGTCGCTCAGAGCGGACA	(8) CAGAARGTGGTGGCTGGTG Human- CAGAAAGTGGTGGCTGGTG Polar- CAGAAGGTGGTGGCTGGTG

Figure S3.4. sgRNA design and construction.

Oligonucleotides designed for hemoglobin depletion from full-length cDNA. Oligos were chosen to deplete Hemoglobin mRNA transcripts from Human and Polar Bear whole blood. A) To generate sgRNAs a template free PCR was performed to anneal the tracrRNA oligo to an oligo containing the target sequence to generate a full-length oligo. The full-length oligos were converted into sgRNA templates using in-vitro transcription. B) Target oligos used for generating sgRNAs. Degenerate bases are highlighted in grey. (r) indicates reverse orientation

Sample Name	Sequencer	Run Type	Library Prep	Read Number	Median Accuracy
PB3_depleted	Illumina HiSeqX	2X151	Smart-seq2	22322746	N/A
PB19_depleted	Illumina HiSeqX	2X151	Smart-seq2	19418607	N/A
PB21_depleted	Illumina HiSeqX	2X151	Smart-seq2	22467660	N/A
PB3_undepleted	Illumina HiSeqX	2X151	Smart-seq2	58088942	N/A
PB19_undepleted	Illumina HiSeqX	2X151	Smart-seq2	105936701	N/A
PB21_undepleted	Illumina HiSeqX	2X151	Smart-seq2	63096050	N/A
PB19_undepleted	ONT MinION	9.4.1/LSK109	R2C2	5313	93%
PB3_depleted_R1	ONT MinION	9.4.1/LSK109	R2C2	390526	93%
PB3_depleted_R2	ONT MinION	9.4.1/LSK109	R2C2	1691780	94%
PB19_depleted_R1	ONT MinION	9.4.1/LSK109	R2C2	59097	93%
PB19_depleted_R2	ONT MinION	9.4.1/LSK109	R2C2	866087	97.5%
PB21_depleted	ONT MinION	9.4.1/LSK109	R2C2	830952	92%

Table S3.1: High-throughput sequencing runs and characteristics
For R2C2/ONT MinION runs, fully processed R2C2 read numbers and median accuracies are given. Some R2C2/ONT MinION runs were multiplexed, sometimes with samples unrelated to this study. Samples PB19_depleted_R2 and PB3_depleted_R2 represent the current output of the R2C2/ONT MinION combination.

RT
>Oligo-dT-smartseq2
/5Me-
isodC/AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
>TSO_Smartseq2
AAGCAGTGGTATCAACGCAGAGTACATrGrGrG

Primers for amplifying cDNA
>ISPCR
AAGCAGTGGTATCAACGCAGAGT

Tn5 Oligos (Smart-seq2 library prep)
>Tn5ME-R
[phos]CTGTCTCTTATACACATCT
>Tn5ME-A
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
>Tn5ME-B
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Primers for amplifying Tn5 Product
>Nextera_Primer_A
AATGATACGGCGACCACCGAGATCTACAC [8bp i5 index] TCGTCGGCAGCGTCAGATG
>Nextera_Primer_B
CAAGCAGAAGACGGCATACGAGAT [8bp i7 index] GTCTCGTGGGCTCGGAGATGTGTAT

R2C2 Splint_Oligos
>UMI_Splint_1_Forward_ISPCR (Matches ISPCR Primer)
ACTCTGCGTTGATACCACTGCTTTGAGGCTGATGAGTTCCATANNNNNTATATNNNNNAT
CACTACTTAGTTTTTGATAGCTTCAAGCCAGAGTTGTCTTTTTCTCTTTGCTGGCAGTAAA
AG
>UMI_Splint_1_Reverse_ISPCR (Matches ISPCR Primer)
ACTCTGCGTTGATACCACTGCTTAAAGGGATATTTTCGATCGCNNNNNATATANNNNNTT
AGTGCATTTGATCCTTTTACTCCTCCTAAAGAACAACCTGACCCAGCAAAAGGTACACAA
TACTTTTACTGCCAGCAAAGAG
>UMI_Splint_2_Forward_ISPCR (Matches ISPCR Primer)
ACTCTGCGTTGATACCACTGCTTTGCCGGTTGGGTATCAATAANNNNNTATATNNNNNATT
GCCTTTATTCTATCTACTTAGTTTTGGCGATGTAGTCTACCTATCCTGATGCTGAATAAAG
GC
>UMI_Splint_2_Reverse_ISPCR (Matches ISPCR Primer)
ACTCTGCGTTGATACCACTGCTTAATTAGGTTCTAGGATCACGNNNNNATATANNNNNCT
GCCATCGAAAATTTTTCACCCGTAACAAGAACTTACAACTCTCTGACGCCTATATCATGAA
GGCCTTTATTCAGCATCAGGA

Table S3.2 Oligos used in the Long-DASH

Oligos are shown 5’->3’ and were ordered from Integrated DNA Technologies (IDT). Lower case ‘r’ = RNA bases. Spaces are for visual emphasis only.

Conclusions

The work I present here showcases how we can overcome current limitations for capturing an accurate transcriptome. Short-read RNAseq still suffers from inherent issues of limited read length, making full-length mRNA transcript structure analysis challenging. This is particularly true when analyzing the ends of the molecules (Fig. 0.1) where annotating transcription start-sites and end-sites becomes difficult. In order to define these features a targeted approach must be considered. Thus, I have developed Tn5Prime, a single cell 5' capturing technique capable of analyzing transcription start sites. This approach allows one to define 5' ends of molecules that often get lost in standard RNAseq methods. Identifying features like these globally can help us understand how genes are regulated at the level of transcription, such as identifying cell-type specific transcriptional regulatory networks, where transcription factor binding motifs can be inferred near promoter regions. Elucidating such features can also help predict novel transcripts, bringing more data to the emerging thought that alternative transcription and not alternative splicing are the main drivers of promoting isoform diversity across different tissues (Pal et al. 2011). Although this method was developed as a 5' enrichment tool, it could easily be adapted for a 3' enrichment method as well. By integrating both data types it could offer insight into translational control by identifying regulatory elements within both the 5' and 3' UTRs and their effect on translational outcomes.

Although certain features can be specifically targeted using short-read technology, it is simpler to capture from 'end-to-end' complete transcripts from single cells. Here, I have shown that using a long-read sequencing approach it is possible to quantify and identify isoforms at the single cell level. This study also highlights how much mRNA isoform heterogeneity occurs within surface receptors among a seeming homogenous B cell population. This begs the question of how much surface receptor diversity is there? And how much of our receptor landscape is actually targetable? Beyond increasing our understanding of B cell receptor expression, I believe this method can be a powerful tool for patients undergoing cancer immunotherapies. Interactions between the immune system and cancer cells are dynamic and are always evolving from the initial progression of cancer to the development of metastasis. Cancer specific splicing events are known to alter the receptor landscape, conferring drug/therapy resistance due to antigen loss. The interface between immune cells and cancer cells are dynamic, but it can be manipulated by changing the antigen landscape. Patients suffering from lymphomas may contain a specific clonal population expressing similar isoforms, which could be targeted by immunotherapy. By incorporating this long-read single cell method we could make better genetic predictions in a preclinical setting, increasing the odds of successful treatment.

There have already been major improvements to this method by incorporating R2C2 (Volden et al. 2018), which has improved the read accuracy from the standard 84% to 97.5%. Improving base accuracy is beneficial for observing allele specific transcription, somatic mutations and helps resolve unique isoforms with better precision. Increase in base accuracy also expands the throughput capacity by resolving cellular barcodes, necessary for multiplexing thousands of single cells.

With the introduction of single-molecule sequencing, such as purveyed by PacBio and Oxford Nanopore Technologies (ONT), it has proven advantageous for identifying novel isoforms, long-noncoding RNAs and fusion transcripts. However, sequencing throughput and accuracy still wanes compared to the short-read technology. With the implementation of our R2C2 method we have improved the base accuracy, but throughput still remains a limitation (Table 0.1). To increase the read throughput, I developed Long-DASH to remove unwanted abundant transcripts from the sequencing pool to capture more rare transcripts. This method excludes transcripts that could account for the majority of sequencing reads, greatly improving throughput. By eliminating highly abundant transcripts such as hemoglobin and rRNA (unpublished), Long-DASH serves as an enrichment strategy that captures information about the transcriptome that would have otherwise been missed or require a much greater amount of sequencing reads to be detected.

Future Outlook

Another hurdle to overcome when analyzing the transcriptome is the issue regarding length bias. Length bias is often attributed to how samples are prepared specifically during PCR amplification and library preparation. PCR becomes problematic in that within a few cycles of PCR, longer transcripts tend to be overwhelmed by smaller transcripts, which tend to be picked up more easily during sequencing. This can lead the researcher to believe that these longer transcripts are rarer or simply never expressed in a given sample. One way to get around this is to eliminate PCR. R2C2 is ideal for making it into a PCR-free method due to Phi29's ability to copy transcripts at equal lengths independent of the raw transcript size (Fig. 4.1). This is likely due to Phi29's high processivity. Additionally, Phi29 performs a linear amplification versus exponential amplification which greatly reduces amplification bias. Below is a schematic of how a PCR-free method would be incorporated (Fig.4.2A). Preliminary data shows that transcript length distributions are skewed towards shorter transcripts when PCR is performed compared to the newest R2C2v2 where PCR is not performed (Fig.4.2B). By eliminating PCR from our R2C2 method, this could eliminate false chimeric molecules, PCR errors, or skewed quantification all of which are by-products of PCR.

When RNAseq was first developed in the early 1990's there were great expectations as to what a researcher could accomplish. It was thought that not only would we be able to take inventory of all the RNA species within a sample but the transcript structure would also be known. Short read RNAseq is most adept at interrogating transcriptomes to observe global changes in gene expression but fail to capture structure and all RNA within a single sample. Long read full-length sequencing opens the door to the possibility of not only capturing structural information, but can be quantitative as well. Capturing all RNA species still remains a challenge as most long read technologies struggle with smaller RNAs. However, using the R2C2 approach multiple copies are made into one long concatenated read, making it possible to sequence smaller RNAs as well such as tRNAs. Just like assembling an excellent genome requires different types of methods to be combined, maybe annotating an excellent transcriptome may require different technologies as well. However, it is possible that in the foreseeable future ONT will dramatically increase their sequencing capacity and base accuracy. Since these sequencing errors are fairly systematic this will however be a herculean task. Until then, researchers like myself have been at the forefront at making changes in order to harness the power of the long-read technology to making the best snapshot of transcriptomes.

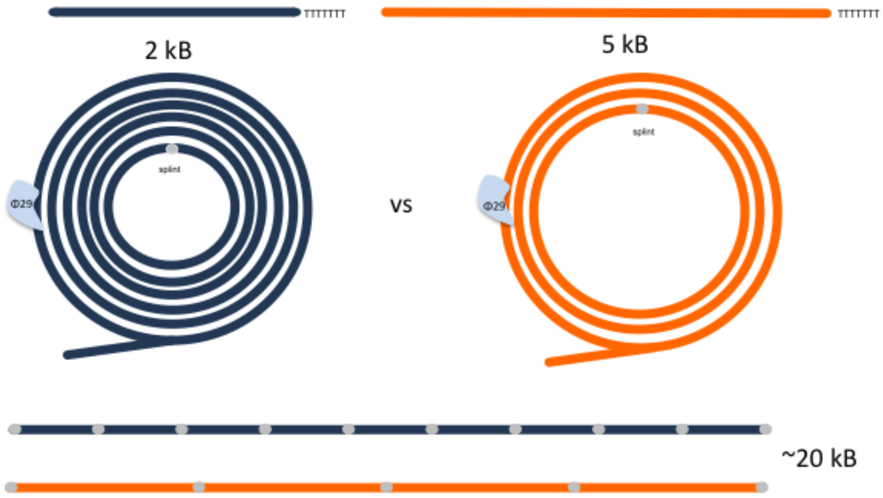


Figure 4.1. Simple schematic of Phi29 amplification. Transcripts of varying sizes are amplified using linear amplification each transcript contains a different number of copies but the raw read length stays the same. Not shown is the branching caused by Phi29 polymerase.

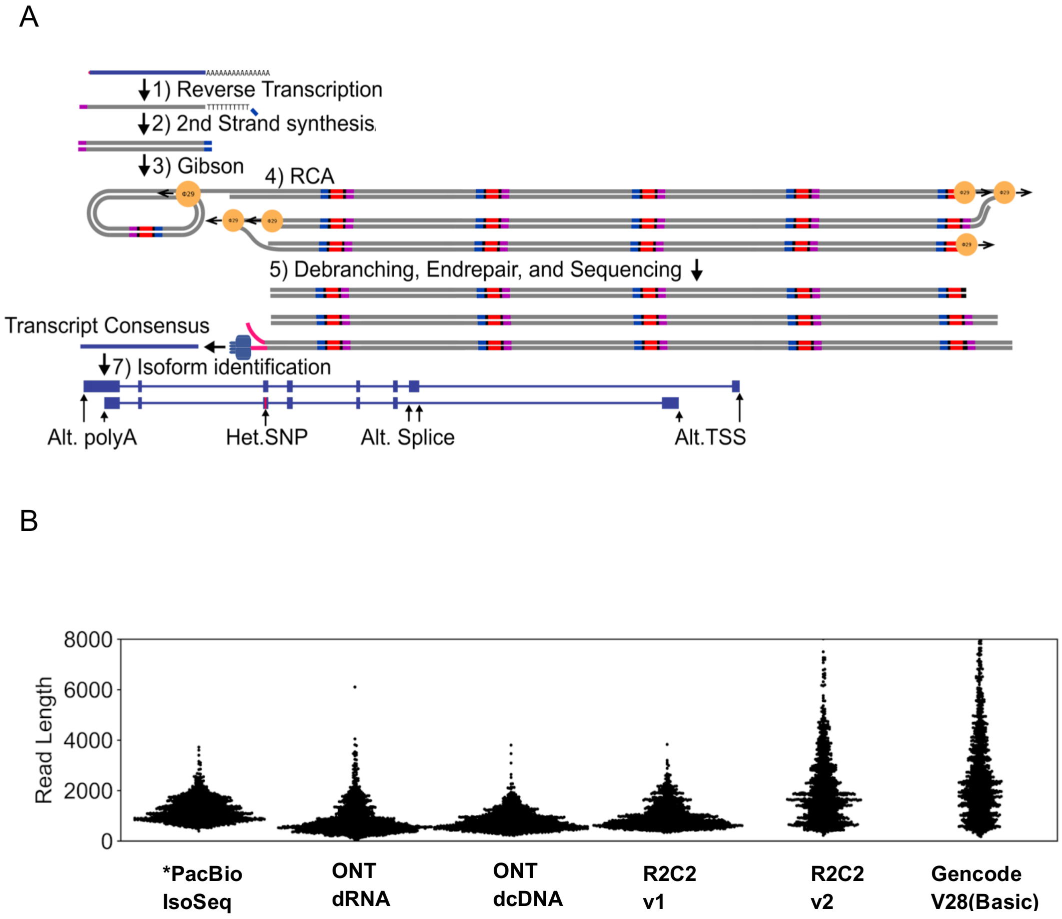


Figure 4.2: PCR-free R2C2 method improves transcript read length.

A) Schematic of PCR-free method. RNA is reverse transcribed and second strand is made. cDNA is then circularized using Gibson Assembly. Rolling circle is performed and the sample is cleaned up and debranched prior to sequencing. B) Swarmplots of length distributions of 1000 randomly sampled PacBio (Tilgner et al. 2014), ONT dRNA, and dcDNA (Workman et al. 2018), R2C2 reads with no adjustments (v1) and R2C2 reads generated from PCR-free method covering the GM12878 (human lymphoblast cell line) transcriptome. These distributions are not representative of the length distribution of the human transcriptome as annotated by GENCODE. *While we show the most recent data set on GM12878 the data provided PacBio technology is several years old and might not be fully representative of current platform performance.

Bibliography

Adams, M. D., M. Dubnick, A. R. Kerlavage, R. Moreno, J. M. Kelley, T. R. Utterback, J. W. Nagle, C. Fields, and J. C. Venter. 1992. “Sequence Identification of 2,375 Human Brain Genes.” *Nature* 355 (6361): 632–34.

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.

Andreadis, Athena. 2005. “Tau Gene Alternative Splicing: Expression Patterns, Regulation and Modulation of Function in Normal Brain and Neurodegenerative Diseases.” *Biochimica et Biophysica Acta* 1739 (2-3): 91–103.

Au, Kin Fai, Vittorio Sebastiano, Pegah Tootoonchi Afshar, Jens Durruthy Durruthy, Lawrence Lee, Brian A. Williams, Harm 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.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (5): 455–77.

Beaudin, Anna E., Scott W. Boyer, and E. Camilla Forsberg. 2014. “Flk2/Flt3 Promotes Both Myeloid and Lymphoid Development by Expanding Non-Self-Renewing

Multipotent Hematopoietic Progenitor Cells." *Experimental Hematology* 42 (3): 218–29.e4.

Beaudin, Anna E., Scott W. Boyer, Jessica Perez-Cunningham, Gloria E. Hernandez, S. Christopher Derderian, Chethan Jujjavarapu, Eric Aaserude, Tippi MacKenzie, and E. Camilla Forsberg. 2016. "A Transient Developmental Hematopoietic Stem Cell Gives Rise to Innate-like B and T Cells." *Cell Stem Cell* In press (September). https://doi.org/10.1016/j.stem.2016.08.013.

Beaudin, Anna E., and E. Camilla Forsberg. 2016. "To B1a or Not to B1a: Do Hematopoietic Stem Cells Contribute to Tissue-Resident Immune Cells?" *Blood* 128 (24): 2765–69.

Bhargava, Vipul, Steven R. Head, Phillip Ordoukhanian, Mark Mercola, and Shankar Subramaniam. 2014. "Technical Variations in Low-Input RNA-Seq Methodologies." *Scientific Reports* 4 (January): 3678.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

Bolisetty, Mohan T., and Karen L. Beemon. 2012. "Splicing of Internal Large Exons Is Defined by Novel Cis-Acting Sequence Elements." *Nucleic Acids Research* 40 (18): 9244–54.

Bolisetty, Mohan T., Gopinath Rajadinakaran, and Brenton R. Graveley. 2015. "Determining Exon Connectivity in Complex mRNAs by Nanopore Sequencing." *Genome Biology* 16 (1): 204.

Borovecki, F., L. Lovrecic, J. Zhou, H. Jeong, F. Then, H. D. Rosas, S. M. Hersch, et al. 2005. “Genome-Wide Expression Profiling of Human Blood Reveals Biomarkers for Huntington’s Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (31): 11023–28.

Brooks, Angela N., Peter S. Choi, Luc de Waal, Tanaz Sharifnia, Marcin Imielinski, Gordon Saksena, Chandra Sekhar Pedamallu, et al. 2014. “A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events.” *PloS One* 9 (1): e87361.

Brown, Tanya M., S. Austin Hammond, Bahar Behsaz, Nik Veldhoen, Inanç Birol, and Caren C. Helbing. 2017. “De Novo Assembly of the Ringed Seal (Pusa Hispida) Blubber Transcriptome: A Tool That Enables Identification of Molecular Health Indicators Associated with PCB Exposure.” *Aquatic Toxicology* 185 (April): 48–57.

Bulletin, User. n.d. “Guidelines for Preparing cDNA Libraries for Isoform Sequencing (Iso-Seq™).” http://www.pacb.com/wp-content/uploads/2015/09/User-Bulletin-Guidelines-for-Preparing-cDNA-Libraries-for-Isoform-Sequencing-Iso-Seq.pdf.

Burgess, Stewart T. G., Andrew Greer, David Frew, Beth Wells, Edward J. Marr, Alasdair J. Nisbet, and John F. Huntley. 2012. “Transcriptomic Analysis of Circulating Leukocytes Reveals Novel Aspects of the Host Systemic Inflammatory Response to Sheep Scab Mites.” *PloS One* 7 (8): e42778.

Busslinger, M., N. Moschonas, and R. A. Flavell. 1981. “Beta + Thalassemia: Aberrant Splicing Results from a Single Point Mutation in an Intron.” *Cell* 27 (2 Pt 1): 289–98.

Byrne, Ashley, Anna E. Beaudin, Hugh E. Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M. DuBois, E. Camilla Forsberg, Mark Akeson, and Christopher Vollmers. 2017. “Nanopore Long-Read RNAseq Reveals Widespread Transcriptional Variation among the Surface Receptors of Individual B Cells.” *Nature Communications* 8 (July): 16027.

Byrne, Ashley, Megan A. Supple, Roger Volden, Kristin L. Laidre, Beth Shapiro, and Christopher Vollmers. 2019. “Depletion of Hemoglobin Transcripts and Long Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (Ursus Maritimus).” *bioRxiv*. https://doi.org/10.1101/527978.

Calame, Kathryn L., Kuo-I Lin, and Chainarong Tunyaplin. 2003. “Regulatory Mechanisms That Determine the Development and Function of Plasma Cells.” *Annual Review of Immunology* 21: 205–30.

Canzar, Stefan, Karlynn E. Neu, Qingming Tang, Patrick C. Wilson, and Aly A. Khan. 2017. “BASIC: BCR Assembly from Single Cells.” *Bioinformatics* 33 (3): 425–27.

Chaisson, Mark J., and Glenn Tesler. 2012. “Mapping Single Molecule Sequencing Reads Using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory.” *BMC Bioinformatics* 13 (September): 238.

Cherf, Gerald M., Kate R. Lieberman, Hytham Rashid, Christopher E. Lam, Kevin Karplus, and Mark Akeson. 2012. “Automated Forward and Reverse Ratcheting of DNA in a Nanopore at 5-Å Precision.” *Nature Biotechnology* 30 (4): 344–48.

Choi, Igseo, Hua Bao, Arun Kommadath, Afshin Hosseini, Xu Sun, Yan Meng, Paul Stothard, et al. 2014. “Increasing Gene Discovery and Coverage Using RNA-Seq of Globin RNA Reduced Porcine Blood Samples.” *BMC Genomics* 15 (November): 954.

Cole, Charles, Ashley Byrne, Anna E. Beaudin, E. Camilla Forsberg, and Christopher Vollmers. 2018. “Tn5Prime, a Tn5 Based 5′ Capture Method for Single Cell RNA-Seq.” *Nucleic Acids Research*, March. https://doi.org/10.1093/nar/gky182.

Cornelison, D. D., and B. J. Wold. 1997. “Single-Cell Analysis of Regulatory Gene Expression in Quiescent and Activated Mouse Skeletal Muscle Satellite Cells.” *Developmental Biology* 191 (2): 270–83.

Cronin, Maureen, Krishna Ghosh, Frank Sistare, John Quackenbush, Vincent Vilker, and Catherine O’Connell. 2004. “Universal RNA Reference Materials for Gene Expression.” *Clinical Chemistry* 50 (8): 1464–71.

Darmanis, Spyros, Steven A. Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M. Shuer, Melanie G. Hayden Gephart, Ben A. Barres, and Stephen R. Quake. 2015. “A Survey of Human Brain Transcriptome Diversity at the Single Cell Level.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (23): 7285–90.

Debey, S., U. Schoenbeck, M. Hellmich, B. S. Gathof, R. Pillai, T. Zander, and J. L. Schultze. 2004. “Comparison of Different Isolation Techniques Prior Gene Expression Profiling of Blood Derived Cells: Impact on Physiological Responses, on Overall Expression and the Role of Different Cell Types.” *The Pharmacogenomics Journal* 4 (3): 193–207.

Dixon, Jesse R., Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T. Le, Galip Gürkan Yardımcı, et al. 2018. “Integrative Detection and Analysis of Structural Variation in Cancer Genomes.” *Nature Genetics* 50 (10): 1388–98.

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.

Du, Lianming, Wujiao Li, Zhenxin Fan, Fujun Shen, Mingyu Yang, Zili Wang, Zuoyi Jian, Rong Hou, Bisong Yue, and Xiuyue Zhang. 2015. “First Insights into the Giant Panda (Ailuropoda Melanoleuca) Blood Transcriptome: A Resource for Novel Gene Loci and Immunogenetics.” *Molecular Ecology Resources* 15 (4): 1001–13.

Edge, Peter, Vineet Bafna, and Vikas Bansal. 2017. “HapCUT2: Robust and Accurate Haplotype Assembly for Diverse Sequencing Technologies.” *Genome Research* 27 (5): 801–12.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, Michiel J. L. de Hoon, Vanja Haberle, et al. 2014. “A Promoter-Level Mammalian Expression Atlas.” *Nature* 507 (7493): 462–70.

Femino, A. M., F. S. Fay, K. Fogarty, and R. H. Singer. 1998. “Visualization of Single RNA Transcripts in Situ.” *Science* 280 (5363): 585–90.

Field, Lori A., Rick M. Jordan, Jennifer A. Hadix, Michael A. Dunn, Craig D. Shriver, Rachel E. Ellsworth, and Darrell L. Ellsworth. 2007. "Functional Identity of Genes Detectable in Expression Profiling Assays Following Globin mRNA Reduction of Peripheral Blood Samples." *Clinical Biochemistry* 40 (7): 499–502.

Flusberg, Benjamin A., Dale R. Webster, Jessica H. Lee, Kevin J. Travers, Eric C. Olivares, Tyson A. Clark, Jonas Korlach, and Stephen W. Turner. 2010. "Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing." *Nature Methods* 7 (6): 461–65.

Fu, Yanfang, Jeffry D. Sander, Deepak Reyon, Vincent M. Cascio, and J. Keith Joung. 2014. "Improving CRISPR-Cas Nuclease Specificity Using Truncated Guide RNAs." *Nature Biotechnology* 32 (3): 279–84.

Garalde, Daniel R., Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, et al. 2018. "Highly Parallel Direct RNA Sequencing on an Array of Nanopores." *Nature Methods* 15 (3): 201–6.

Genomic Resources Development Consortium, David W. Coltman, Corey S. Davis, Nick J. Lunn, René M. Malenfant, and Evan S. Richardson. 2014. "Genomic Resources Notes Accepted 1 August 2013-30 September 2013." *Molecular Ecology Resources* 14 (1): 219.

Gervasoni, Annalisa, Rina M. Monasterio Muñoz, Georg S. Wengler, Anna Rizzi, Alberto Zaniboni, and Ornella Parolini. 2008. "Molecular Signature Detection of Circulating Tumor Cells Using a Panel of Selected Genes." *Cancer Letters* 263 (2): 267–79.

Gierahn, Todd M., Marc H. Wadsworth 2nd, Travis K. Hughes, Bryan D. Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J. Christopher Love, and Alex K. Shalek. 2017. “Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput.” *Nature Methods* 14 (4): 395–98.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52.

Graf, Thomas, and Matthias Stadtfeld. 2008. “Heterogeneity of Embryonic and Adult Stem Cells.” *Cell Stem Cell* 3 (5): 480–83.

Graveley, B. R. 2001. “Alternative Splicing: Increasing Diversity in the Proteomic World.” *Trends in Genetics: TIG* 17 (2): 100–107.

Gupta, Ishaan, Paul G. Collier, Bettina Haase, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, et al. 2018. “Single-Cell Isoform RNA Sequencing Characterizes Isoforms in Thousands of Cerebellar Cells.” *Nature Biotechnology*, October. https://doi.org/10.1038/nbt.4259.

Gu, W., E. D. Crawford, B. D. O’Donovan, M. R. Wilson, E. D. Chow, H. Retallack, and J. L. DeRisi. 2016. “Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to Remove Unwanted High-Abundance Species in Sequencing Libraries and Molecular Counting Applications.” *Genome Biology* 17 (1): 41.

Harbers, Matthias, Sachi Kato, Michiel de Hoon, Yoshihide Hayashizaki, Piero Carninci, and Charles Plessy. 2013. “Comparison of RNA- or LNA-Hybrid Oligonucleotides in

Template-Switching Reactions for High-Speed Sequencing Library Preparation." *BMC Genomics* 14 (September): 665.

Hardwick, Simon A., Wendy Y. Chen, Ted Wong, Ira W. Deveson, James Blackburn, Stacey B. Andersen, Lars K. Nielsen, John S. Mattick, and Tim R. Mercer. 2016. "Spliced Synthetic Genes as Internal Controls in RNA Sequencing Experiments." *Nature Methods* 13 (9): 792–98.

Hargreaves, Adam D., and John F. Mulley. 2015. "Assessing the Utility of the Oxford Nanopore MinION for Snake Venom Gland cDNA Sequencing." *PeerJ* 3 (November): e1441.

Harr, Bettina, and Leslie M. Turner. 2010. "Genome-Wide Analysis of Alternative Splicing Evolution among Mus Subspecies." *Molecular Ecology* 19 Suppl 1 (March): 228–39.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project." *Genome Research* 22 (9): 1760–74.

Hernández-Fernández, Javier, Andrés Pinzón, and Leonardo Mariño-Ramírez. 2017. "De Novo Transcriptome Assembly of Loggerhead Sea Turtle Nesting of the Colombian Caribbean." *Genomics Data* 13 (September): 18–20.

Huang, Zixia, Aurore Gallot, Nga T. Lao, Sébastien J. Puechmaille, Nicole M. Foley, David Jebb, Michaël Bekaert, and Emma C. Teeling. 2016. "A Nonlethal Sampling

Method to Obtain, Generate and Assemble Whole Blood Transcriptomes from Small, Wild Mammals.” *Molecular Ecology Resources* 16 (1): 150–62.

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90–95.

Ilagan, Janine O., Aravind Ramakrishnan, Brian Hayes, Michele E. Murphy, Ahmad S. Zebari, Philip Bradley, and Robert K. Bradley. 2015. “U2AF1 Mutations Alter Splice Site Recognition in Hematological Malignancies.” *Genome Research* 25 (1): 14–26.

Irish, Jonathan M., Nikesh Kotecha, and Garry P. Nolan. 2006. “Mapping Normal and Cancer Cell Signalling Networks: Towards Single-Cell Proteomics.” *Nature Reviews. Cancer* 6 (2): 146–55.

Islam, K. B., B. Baskin, B. Christensson, L. Hammarström, and C. I. Smith. 1994. “In Vivo Expression of Human Immunoglobulin Germ-Line mRNA in Normal and in Immunodeficient Individuals.” *Clinical and Experimental Immunology* 95 (1): 3–9.

Islam, Saiful, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. 2011. “Characterization of the Single-Cell Transcriptional Landscape by Highly Multiplex RNA-Seq.” *Genome Research* 21 (7): 1160–67.

Islam, Saiful, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. 2014. “Quantitative Single-Cell RNA-Seq with Unique Molecular Identifiers.” *Nature Methods* 11 (2): 163–66.

Jain, Miten, Ian T. Fiddes, Karen H. Miga, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2015. “Improved Data Analysis for the MinION Nanopore Sequencer.” *Nature Methods* 12 (4): 351–56.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology*, January. https://doi.org/10.1038/nbt.4060.

Jaitin, Diego Adhemar, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, et al. 2014. “Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types.” *Science* 343 (6172): 776–79.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.” *Science* 337 (6096): 816–21.

Jones, Eric, Travis Oliphant, and Pearu Peterson. 2001--. “{SciPy}: Open Source Scientific Tools for {Python}.” http://www.scipy.org.

Jones, Eric, Travis Oliphant, Pearu Peterson, and Others. n.d. “{SciPy}: Open Source Scientific Tools for {Python}.” http://www.scipy.org/.

Kalsotra, Auinash, Xinshu Xiao, Amanda J. Ward, John C. Castle, Jason M. Johnson, Christopher B. Burge, and Thomas A. Cooper. 2008. “A Postnatal Switch of CELF and MBNL Proteins Reprograms Alternative Splicing in the Developing Heart.”

Proceedings of the National Academy of Sciences of the United States of America 105 (51): 20333–38.

Kanitz, Alexander, Foivos Gypas, Andreas J. Gruber, Andreas R. Gruber, Georges Martin, and Mihaela Zavolan. 2015. “Comparative Assessment of Methods for the Computational Inference of Transcript Isoform Abundance from RNA-Seq Data.” *Genome Biology* 16 (July): 150.

Kent, W. James. 2002. “BLAT—The BLAST-Like Alignment Tool.” *Genome Research* 12 (4): 656–64.

Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6): 996–1006.

Khudyakov, J. I., C. D. Champagne, L. M. Meneghetti, and D. E. Crocker. 2017. “Blubber Transcriptome Response to Acute Stress Axis Activation Involves Transient Changes in Adipogenesis and Lipolysis in a Fasting-Adapted Marine Mammal.” *Scientific Reports* 7 (February): 42110.

Kilianski, Andy, Jamie L. Haas, Elizabeth J. Corriveau, Alvin T. Liem, Kristen L. Willis, Dana R. Kadavy, C. Nicole Rosenzweig, and Samuel S. Minot. 2015. “Bacterial and Viral Identification and Differentiation by Amplicon Sequencing on the MinION Nanopore Sequencer.” *GigaScience* 4 (March): 12.

Kim, Jeong Kyu, Kwang Hwa Jung, Ji Heon Noh, Jung Woo Eun, Hyun Jin Bae, Hong Jian Xie, Ja-June Jang, et al. 2011. “Identification of Characteristic Molecular

Signature for Volatile Organic Compounds in Peripheral Blood of Rat." *Toxicology and Applied Pharmacology* 250 (2): 162–69.

Koren, Sergey, Michael C. Schatz, Brian P. Walenz, Jeffrey Martin, Jason T. Howard, Ganeshkumar Ganapathy, Zhong Wang, et al. 2012. "Hybrid Error Correction and de Novo Assembly of Single-Molecule Sequencing Reads." *Nature Biotechnology* 30 (7): 693–700.

Krawczak, M., J. Reiss, and D. N. Cooper. 1992. "The Mutational Spectrum of Single Base-Pair Substitutions in mRNA Splice Junctions of Human Genes: Causes and Consequences." *Human Genetics* 90 (1-2): 41–54.

Lamson, G., and M. E. Koshland. 1984. "Changes in J Chain and Mu Chain RNA Expression as a Function of B Cell Differentiation." *The Journal of Experimental Medicine* 160 (3): 877–92.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.

Lee, Christopher, Catherine Grasso, and Mark F. Sharlow. 2002. "Multiple Sequence Alignment Using Partial Order Graphs." *Bioinformatics* 18 (3): 452–64.

Lefranc, Marie-Paule, Véronique Giudicelli, Chantal Ginestoux, Nathalie Bosc, Géraldine Folch, Delphine Guiraudou, Joumana Jabado-Michaloud, et al. 2004. "IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics." *In Silico Biology* 4 (1): 17–29.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30.

Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12 (August): 323.

Li, Chenhao, Kern Rei Chng, Esther Jia Hui Boey, Amanda Hui Qi Ng, Andreas Wilm, and Niranjan Nagarajan. 2016. “INC-Seq: Accurate Single Molecule Reads Using Nanopore Sequencing.” *GigaScience* 5 (1): 34.

Liew, Choong-Chin, Jun Ma, Hong-Chang Tang, Run Zheng, and Adam A. Dempsey. 2006. “The Peripheral Blood Transcriptome Dynamically Reflects System Wide Biology: A Potential Diagnostic Tool.” *The Journal of Laboratory and Clinical Medicine* 147 (3): 126–32.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv Preprint arXiv:1303. 3997*. https://arxiv.org/abs/1303.3997.

Li, Heng. 2017. “Minimap2: Fast Pairwise Alignment for Long Nucleotide Sequences.” *ArXiv E-Prints* 2017. https://pdfs.semanticscholar.org/a703/88011f2995783e159dc21a62905753a6af44.pdf.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Lindenbaum, Pierre. 2015. “JVarkit: Java-Based Utilities for Bioinformatics.” *Figshare*, May. https://doi.org/10.6084/m9.figshare.1425030.v1.

Li, Ruiqiang, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, et al. 2010. “The Sequence and de Novo Assembly of the Giant Panda Genome.” *Nature* 463 (7279): 311–17.

Liu, Shiping, Eline D. Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou, et al. 2014. “Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears.” *Cell* 157 (4): 785–94.

Li, Weizhong, Andrew Cowley, Mahmut Uludag, Tamer Gur, Hamish McWilliam, Silvano Squizzato, Young Mi Park, Nicola Buso, and Rodrigo Lopez. 2015. “The EMBL-EBI Bioinformatics Web and Programmatic Tools Framework.” *Nucleic Acids Research* 43 (W1): W580–84.

Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, et al. 1996. “Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays.” *Nature Biotechnology* 14 (13): 1675–80.

Loman, Nicholas J., Joshua Quick, and Jared T. Simpson. 2015. “A Complete Bacterial Genome Assembled de Novo Using Only Nanopore Sequencing Data.” *Nature Methods* 12 (8): 733–35.

Loman, Nicholas J., and Aaron R. Quinlan. 2014. “Poretools: A Toolkit for Analyzing Nanopore Sequence Data.” *Bioinformatics* 30 (23): 3399–3401.

Lun, Aaron T. L., Karsten Bach, and John C. Marioni. 2016. “Pooling across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts.” *Genome Biology* 17 (April): 75.

Lv, Jianliang, Yaozhong Ding, Xinsheng Liu, Li Pan, Zhongwang Zhang, Peng Zhou, Yongguang Zhang, and Yonghao Hu. 2018. “Gene Expression Analysis of Porcine Whole Blood Cells Infected with Foot-and-Mouth Disease Virus Using High-Throughput Sequencing Technology.” *PloS One* 13 (7): e0200081.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9 (Nov): 2579–2605.

Macaulay, Iain C., Wilfried Haerty, Parveen Kumar, Yang I. Li, Tim Xiaoming Hu, Mabel J. Teng, Mubeen Goolam, et al. 2015. “G&T-Seq: Parallel Sequencing of Single-Cell Genomes and Transcriptomes.” *Nature Methods* 12 (6): 519–22.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.

Mastrokolias, Anastasios, Johan T. den Dunnen, Gertjan B. van Ommen, Peter A. C. ’t Hoen, and Willeke M. C. van Roon-Mom. 2012. “Increased Sensitivity of next Generation Sequencing-Based Expression Profiling after Globin Reduction in Human Blood RNA.” *BMC Genomics* 13 (January): 28.

McLoughlin, Kirsten E., Nicolas C. Nalpas, Kévin Rue-Albrecht, John A. Browne, David A. Magee, Kate E. Killick, Stephen D. E. Park, et al. 2014. “RNA-Seq Transcriptional Profiling of Peripheral Blood Leukocytes from Cattle Infected with Mycobacterium Bovis.” *Frontiers in Immunology* 5 (August): 396.

McWilliam, Hamish, Weizhong Li, Mahmut Uludag, Silvano Squizzato, Young Mi Park, Nicola Buso, Andrew Peter Cowley, and Rodrigo Lopez. 2013. “Analysis Tool Web Services from the EMBL-EBI.” *Nucleic Acids Research* 41 (Web Server issue): W597–600.

Medaka. n.d. Github. Accessed February 12, 2019. https://github.com/nanoporetech/medaka.

Millman, K. Jarrod, and Michael Aivazis. 2011. “Python for Scientists and Engineers.” *Computing in Science & Engineering* 13 (2): 9–12.

Minnich, Martina, Hiromi Tagoh, Peter Bönelt, Elin Axelsson, Maria Fischer, Beatriz Cebolla, Alexander Tarakhovsky, Stephen L. Nutt, Markus Jaritz, and Meinrad Busslinger. 2016. “Multifunctional Role of the Transcription Factor Blimp-1 in Coordinating Plasma Cell Differentiation.” *Nature Immunology* 17 (3): 331–43.

Modrek, Barmak, and Christopher Lee. 2002. “A Genomic View of Alternative Splicing.” *Nature Genetics* 30 (1): 13–19.

Morey, Jeanine S., Marion G. Neely, Denise Lunardi, Paul E. Anderson, Lori H. Schwacke, Michelle Campbell, and Frances M. Van Dolah. 2016. “RNA-Seq Analysis of Seasonal and Individual Variation in Blood Transcriptomes of Healthy Managed Bottlenose Dolphins.” *BMC Genomics* 17 (September): 720.

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.

Mudge, Jonathan M., and Jennifer Harrow. 2015. “Creating Reference Gene Annotation for the Mouse C57BL6/J Genome Assembly.” *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 26 (9-10): 366–78.

Mure, Ludovic S., Hiep D. Le, Giorgia Benegiamo, Max W. Chang, Luis Rios, Ngalla Jillani, Maina Ngotho, et al. 2018. “Diurnal Transcriptome Atlas of a Primate across Major Neural and Peripheral Tissues.” *Science* 359 (6381). https://doi.org/10.1126/science.aao0318.

Nutt, Stephen L., Philip D. Hodgkin, David M. Tarlinton, and Lynn M. Corcoran. 2015. “The Generation of Antibody-Secreting Plasma Cells.” *Nature Reviews. Immunology* 15 (3): 160–71.

Oikonomopoulos, Spyros, Yu Chang Wang, Haig Djambazian, Dunarel Badescu, and Jiannis Ragoussis. 2016. “Benchmarking of the Oxford Nanopore MinION Sequencing for Quantitative and Qualitative Assessment of cDNA Populations.” *Scientific Reports* 6 (August): 31602.

Oliphant, Travis E. 2007. “Python for Scientific Computing.” *Computing in Science & Engineering* 9 (3): 10–20.

Pal, Sharmistha, Ravi Gupta, Hyunsoo Kim, Priyankara Wickramasinghe, Valérie Baubet, Louise C. Showe, Nadia Dahmane, and Ramana V. Davuluri. 2011. “Alternative

Transcription Exceeds Alternative Splicing in Generating the Transcriptome Diversity of Cerebellar Development." *Genome Research* 21 (8): 1260–72.

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.

Park, Eddie, Brian Williams, Barbara J. Wold, and Ali Mortazavi. 2012. "RNA Editing in the Human ENCODE RNA-Seq Data." *Genome Research* 22 (9): 1626–33.

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.

Picelli, Simone, Åsa K. Björklund, Omid R. Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2013. "Smart-seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells." *Nature Methods* 10 (11): 1096–98.

Picelli, Simone, Asa K. Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2014. "Tn5 Transposase and Tagmentation Procedures for Massively Scaled Sequencing Projects." *Genome Research* 24 (12): 2033–40.

Picelli, Simone, Omid R. Faridani, Asa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. "Full-Length RNA-Seq from Single Cells Using Smart-seq2." *Nature Protocols* 9 (1): 171–81.

Pruitt, Kim D., Garth R. Brown, Susan M. Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M. Farrell, et al. 2014. "RefSeq: An Update on

Mammalian Reference Sequences." *Nucleic Acids Research* 42 (Database issue): D756–63.

Putnam, Nicholas H., Brendan L. O'Connell, Jonathan C. Stites, Brandon J. Rice, Marco Blanchette, Robert Calef, Christopher J. Troll, et al. 2016. "Chromosome-Scale Shotgun Assembly Using an in Vitro Method for Long-Range Linkage." *Genome Research* 26 (3): 342–50.

Raj, Arjun, Patrick van den Bogaard, Scott A. Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. 2008. "Imaging Individual mRNA Molecules Using Multiple Singly Labeled Probes." *Nature Methods* 5 (10): 877–79.

Rasband, Wayne S. 2011. "Imagej, Us National Institutes of Health, Bethesda, Maryland, Usa." *Http://imagej. Nih. Gov/ij/*. https://ci.nii.ac.jp/naid/10030139275/.

Ren, Xingjie, Zhihao Yang, Jiang Xu, Jin Sun, Decai Mao, Yanhui Hu, Su-Juan Yang, et al. 2014. "Enhanced Specificity and Efficiency of the CRISPR/Cas9 System with Optimized sgRNA Parameters in Drosophila." *Cell Reports* 9 (3): 1151–62.

Ruan, Xiaoan, and Yijun Ruan. 2011. "RNA-PET: Full-Length Transcript Analysis Using 5′- and 3′-Paired-End Tag Next-Generation Sequencing." In *Tag-Based Next Generation Sequencing*, 73–90. Wiley-VCH Verlag GmbH & Co. KGaA.

Salimullah, Md, Mizuho Sakai, Sakai Mizuho, Charles Plessy, and Piero Carninci. 2011. "NanoCAGE: A High-Resolution Technique to Discover and Interrogate Cell Transcriptomes." *Cold Spring Harbor Protocols* 2011 (1): db.prot5559.

Salzberg, Steven L. 2019. "Next-Generation Genome Annotation: We Still Struggle to Get It Right." *Genome Biology* 20 (1): 92.

Salzberg, Steven L., Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, et al. 2012. "GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms." *Genome Research* 22 (3): 557–67.

Seabold, Skipper, and Josef Perktold. n.d. "Statsmodels: Econometric and Statistical Modeling with Python." http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf.

Shalek, Alex K., Rahul Satija, Xian Adiconis, Rona S. Gertner, Jellert T. Gaublomme, Raktima Raychowdhury, Schraga Schwartz, et al. 2013. "Single-Cell Transcriptomics Reveals Bimodality in Expression and Splicing in Immune Cells." *Nature* 498 (7453): 236–40.

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.

Shin, Heesun, Casey P. Shannon, Nick Fishbane, Jian Ruan, Mi Zhou, Robert Balshaw, Janet E. Wilson-McManus, et al. 2014. "Variation in RNA-Seq Transcriptome Profiles of Peripheral Whole Blood from Healthy Individuals with and without Globin Depletion." *PloS One* 9 (3): e91041.

Shiraki, Toshiyuki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, et al. 2003. "Cap Analysis Gene Expression for High-Throughput Analysis of Transcriptional Starting Point and Identification of Promoter Usage." *Proceedings of the National Academy of Sciences of the United States of America* 100 (26): 15776–81.

Shi, Xiaoli, Danny W-K Ng, Changqing Zhang, Luca Comai, Wenxue Ye, and Z. Jeffrey Chen. 2012. “Cis- and Trans-Regulatory Divergence between Progenitor Species Determines Gene-Expression Novelty in Arabidopsis Allopolyploids.” *Nature Communications* 3 (July): 950.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (October): 539.

Sims, David, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. 2014. “Sequencing Depth and Coverage: Key Considerations in Genomic Analyses.” *Nature Reviews. Genetics* 15 (2): 121–32.

Sinha, Rahul, Geoff Stanley, Gunsagar Singh Gulati, Camille Ezran, Kyle Joseph Travaglini, Eric Wei, Charles Kwok Fai Chan, et al. 2017. “Index Switching Causes ‘Spreading-Of-Signal’ Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing.” *bioRxiv*. https://doi.org/10.1101/125724.

Smith-Berdan, Stephanie, Andrew Nguyen, Matthew A. Hong, and E. Camilla Forsberg. 2015. “ROBO4-Mediated Vascular Integrity Regulates the Directionality of Hematopoietic Stem Cell Trafficking.” *Stem Cell Reports* 4 (2): 255–68.

Stamm, Stefan, Shani Ben-Ari, Ilona Rafalska, Yesheng Tang, Zhaiyi Zhang, Debra Toiber, T. A. Thanaraj, and Hermona Soreq. 2005. “Function of Alternative Splicing.” *Gene* 344 (January): 1–20.

Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern. 2004. “AUGUSTUS: A Web Server for Gene Finding in Eukaryotes.” *Nucleic Acids Research* 32 (Web Server issue): W309–12.

Sugnet, C. W., W. J. Kent, M. Ares Jr, and D. Haussler. 2004. “Transcriptome and Genome Conservation of Alternative Splicing Events in Humans and Mice.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 66–77.

Supple, Megan A., and Beth Shapiro. 2018. “Conservation of Biodiversity in the Genomics Era.” *Genome Biology* 19 (1): 131.

Tang, Fuchou, Kaiqin Lao, and M. Azim Surani. 2011. “Development and Applications of Single-Cell Transcriptome Analysis.” *Nature Methods* 8 (4 Suppl): S6–11.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector Del Risco, Marc Ferrell, et al. 2018. “SQANTI: Extensive Characterization of Long-Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification.” *Genome Research*, February. https://doi.org/10.1101/gr.222976.117.

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

Tilgner, Hagen, Fabian Grubert, Donald Sharon, and Michael P. Snyder. 2014. “Defining a Personal, Allele-Specific, and Single-Molecule Long-Read Transcriptome.”

Proceedings of the National Academy of Sciences of the United States of America 111 (27): 9869–74.

Tilgner, Hagen, Fereshteh Jahanbani, Tim Blauwkamp, Ali Moshrefi, Erich Jaeger, Feng Chen, Itamar Harel, Carlos D. Bustamante, Morten Rasmussen, and Michael P. Snyder. 2015. "Comprehensive Transcriptome Analysis Using Synthetic Long-Read Sequencing Reveals Molecular Co-Association of Distant Splicing Events." *Nature Biotechnology* 33 (7): 736–42.

Tilgner, Hagen, Fereshteh Jahanbani, Ishaan Gupta, Paul Collier, Eric Wei, Morten Rasmussen, and Michael Snyder. 2017. "Microfluidic Isoform Sequencing Shows Widespread Splicing Coordination in the Human Transcriptome." *Genome Research*, December. https://doi.org/10.1101/gr.230516.117.

Tilgner, Hagen, Debasish Raha, Lukas Habegger, Mohammed Mohiuddin, Mark Gerstein, and Michael Snyder. 2013. "Accurate Identification and Analysis of Human mRNA Isoforms Using Deep Long Read Sequencing." *G3* 3 (3): 387–97.

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.

Treutlein, Barbara, Doug G. Brownfield, Angela R. Wu, Norma F. Neff, Gary L. Mantalas, F. Hernan Espinoza, Tushar J. Desai, Mark A. Krasnow, and Stephen R.

Quake. 2014. “Reconstructing Lineage Hierarchies of the Distal Lung Epithelium Using Single-Cell RNA-Seq.” *Nature* 509 (7500): 371–75.

Treutlein, Barbara, Ozgun Gokce, Stephen R. Quake, and Thomas C. Südhof. 2014. “Cartography of Neurexin Alternative Splicing Mapped by Single-Molecule Long-Read mRNA Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (13): E1291–99.

Ugarte, Fernando, Rebekah Sousae, Bertrand Cinquin, Eric W. Martin, Jana Krietsch, Gabriela Sanchez, Margaux Inman, et al. 2015. “Progressive Chromatin Condensation and H3K9 Methylation Regulate the Differentiation of Embryonic and Hematopoietic Stem Cells.” *Stem Cell Reports* 5 (5): 728–40.

Ungaro, Arnaud, Nicolas Pech, Jean-François Martin, R. J. Scott McCairns, Jean-Philippe Mévy, Rémi Chappaz, and André Gilles. 2017. “Challenges and Advances for Transcriptome Assembly in Non-Model Species.” *PloS One* 12 (9): e0185020.

Valk, Peter J. M., Roel G. W. Verhaak, M. Antoinette Beijen, Claudia A. J. Erpelinck, Sahar Barjesteh van Waalwijk van Doorn-Khosrovani, Judith M. Boer, H. Berna Beverloo, et al. 2004. “Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia.” *The New England Journal of Medicine* 350 (16): 1617–28.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27 (5): 737–46.

Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler. 1995. “Serial Analysis of Gene Expression.” *Science* 270 (5235): 484–87.

Volden, Roger, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard E. Green, and Christopher Vollmers. 2018. "Improving Nanopore Read Accuracy with the R2C2 Method Enables the Sequencing of Highly Multiplexed Full-Length Single-Cell cDNA." *Proceedings of the National Academy of Sciences of the United States of America*, September. https://doi.org/10.1073/pnas.1806447115.

Vollmers, Christopher, Lolita Penland, Jad N. Kanbar, and Stephen R. Quake. 2015. "Novel Exons and Splice Variants in the Human Antibody Heavy Chain Identified by Single Cell and Single Molecule Sequencing." *PloS One* 10 (1): e0117050.

Walt, Stéfan van der, S. Chris Colbert, and Gaël Varoquaux. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science & Engineering* 13 (2): 22–30.

Wang, Bo, Elizabeth Tseng, Michael Regulski, Tyson A. Clark, Ting Hon, Yinping Jiao, Zhenyuan Lu, Andrew Olson, Joshua C. Stein, and Doreen Ware. 2016. "Unveiling the Complexity of the Maize Transcriptome by Single-Molecule Long-Read Sequencing." *Nature Communications* 7 (June): 11708.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. 2008. "Alternative Isoform Regulation in Human Tissue Transcriptomes." *Nature* 456 (7221): 470–76.

Warren, Luigi, David Bryder, Irving L. Weissman, and Stephen R. Quake. 2006. "Transcription Factor Profiling in Individual Hematopoietic Progenitors by Digital

RT-PCR." *Proceedings of the National Academy of Sciences of the United States of America* 103 (47): 17807–12.

Watson, Hannah, Elin Videvall, Martin N. Andersson, and Caroline Isaksson. 2017. "Transcriptome Analysis of a Wild Bird Reveals Physiological Responses to the Urban Environment." *Scientific Reports* 7 (March): 44180.

Weirather, Jason L., Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. 2017. "Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis." *F1000Research* 6 (February): 100.

Welch, Joshua D., Yin Hu, and Jan F. Prins. 2016. "Robust Detection of Alternative Splicing in a Population of Single Cells." *Nucleic Acids Research* 44 (8): e73.

Workman, Rachael E., Alison Tang, Paul S. Tang, Miten Jain, John R. Tyson, Philip C. Zuzarte, Timothy Gilpatrick, et al. 2018. "Nanopore Native RNA Sequencing of a Human poly(A) Transcriptome." *bioRxiv*. https://doi.org/10.1101/459529.

Wrammert, Jens, Kenneth Smith, Joe Miller, William A. Langley, Kenneth Kokko, Christian Larsen, Nai-Ying Zheng, et al. 2008. "Rapid Cloning of High-Affinity Human Monoclonal Antibodies against Influenza Virus." *Nature* 453 (7195): 667–71.

Wu, Angela R., Norma F. Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E. Rothenberg, Francis M. Mburu, et al. 2014. "Quantitative Assessment of Single-Cell RNA-Sequencing Methods." *Nature Methods* 11 (1): 41–46.

Wu, Thomas D., and Colin K. Watanabe. 2005. “GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.” *Bioinformatics* 21 (9): 1859–75.

Wu, Yee Ling, Michael J. T. Stubbington, Maria Daly, Sarah A. Teichmann, and Cristina Rada. 2017. “Intrinsic Transcriptional Heterogeneity in B Cells Controls Early Class Switching to IgE.” *The Journal of Experimental Medicine* 214 (1): 183–96.

Ye, Jian, Ning Ma, Thomas L. Madden, and James M. Ostell. 2013. “IgBLAST: An Immunoglobulin Variable Domain Sequence Analysis Tool.” *Nucleic Acids Research* 41 (Web Server issue): W34–40.

Zhang, Xiaochang, Ming Hui Chen, Xuebing Wu, Andrew Kodani, Jean Fan, Ryan Doan, Manabu Ozawa, et al. 2016. “Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex.” *Cell* 166 (5): 1147–62.e15.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January): 14049.

Zhu, Xiaocui, Rebecca Hart, Mi Sook Chang, Jong-Woo Kim, Sun Young Lee, Yun Anna Cao, Dennis Mock, et al. 2004. “Analysis of the Major Patterns of B Cell Gene Expression Changes in Response to Short-Term Stimulation with 33 Single Ligands.” *Journal of Immunology* 173 (12): 7141–49.

Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and

Wolfgang Enard. 2017. “Comparative Analysis of Single-Cell RNA Sequencing Methods.” *Molecular Cell* 65 (4): 631–43.e4.