University of California
Santa Cruz

Antibody Genetics

A dissertation submitted in partial satisfaction of
The requirements for the degree of

DOCTOR OF PHILOSOPHY
IN
BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

BY

Charles K. Cole

September 2019

The Dissertation of Charles K. Cole is approved:

_____
Professor Christopher Vollmers, chair

_____
Professor Angela Brooks

_____
Professor Melissa Jurica

_____
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

Table of figures

Antibody Genetics
Charles K. Cole
Abstract

Antibodies are created through a unique mechanism whereby different gene segments at the antibody loci are semi-randomly recombined to form functional antibody genes. This process allows that adaptive immune system to generate protection against a near-infinite array of pathogens, however it presents unique challenges from a sequencing prospective. I present three projects addressing problems facing the repertoire sequencing community including a method for full-length antibody transcript sequencing, a method for heavy and light chain pairing in single B cells, and a method for generating repertoires as a byproduct of a general polyA RNA sequencing protocol.

Acknowledgements

I would like to acknowledge the help I received from my advisor Christopher Vollmers. He has been an invaluable source of advice over the last five years and, in fact, has been the primary architect of the projects I have pursued. He has inspired me with his dedication to science and teaching. I have been extremely fortunate to have him as a mentor and I will miss him dearly. I would like to acknowledge my labmate Ashley Byrne. Her skills as a molecular biologist are without peer and she is the primary reason we were ever able to sequence single cells in the first place. She has been a wonderful companion throughout my time at UCSC and I wish her the best of luck at her new job at the Chan Zuckerberg Biohub. I would also like to recognize the contributions of Roger Volden, who can claim most of the credit for the development of the R2C2 method. He has been a delight to work with and I expect great things from him in the future. And lastly, I would like to acknowledge the help I received from my mother, father, and sister who have been a never-ending source of moral support.

Antibody Genetics

The purpose of the Immune system is to allow the body to discriminate between self and non-self. This enables the body to detect and neutralize viruses, bacteria, fungi, toxins and, in some situations, even cancer. In Vertebrates, this system can be divided into two mutually supporting but conceptually distinct parts: innate and adaptive immunity. The innate immune system is genetically encoded in the germline and operates by recognizing evolutionarily-conserved patterns in biomolecules which indicate the presence of harmful organisms. The adaptive immune system must learn to recognize pathogens through patterns in protein structure and is, on the whole, slower than the innate immune system, taking several days for the response to occur as opposed to several hours. However, once pathogens are recognized, the adaptive immune system can retain life-long memory of them and provide quick and effective response to future infections.

The adaptive Immune system can be thought of as a chain of responses which start when dendritic cells or macrophages - professional antigen presenting cells engulf the pathogen-associated proteins. These proteins are then fragmented into smaller peptides and presented on the cell surface as part of the Major Histocompatibility Complex (MHC2). These peptide fragments are called antigens and, when presented with the correct co-stimulatory factors and in the correct context, will initiate an immune response by activating CD4+ and CD8+ T cells. Those possessing a T cell receptor (TCR) specific to one of the antigens will continue to proliferate and drive further action by the immune system. CD8+ T cells,

also known as Cytotoxic CD8+ T cells, serve an important role by migrating to the site of infection and killing infected cells. CD4+ T cells, on the other hand support immune response in two important capacities. In their first capacity, they produce cytokines and chemokines which attract CD8+ T cells, Natural Killer cells and other immune cells to the site of infection. In their second capacity they initiate the B cell response by activating B cells in the germinal centers of the lymph nodes and spleen.

A Brief History of Antibody Genetics

The precise moment that B cells were discovered is somewhat debatable because antibodies and the B cell response were discovered before the cell population was isolated. In 1908 Paul Ehrlich received the Nobel prize for the discoveries he made in the field of immunology and specifically for his work on vaccine development. In his work he characterized the production and interaction between “toxins” and substances capable of neutralizing those toxins which he referred to as “antibodies”. He noted that antibodies effective against one toxin would be ineffective against another. From these results he developed the “lock and key” theory of antibody specificity, wherein each antibody is specific to a particular antigen. In 1958 G. J. V. Nossal published a set of experiments where the individual lymph node cells from rats immunized with one or both of two strains of salmonella were isolated in microdroplets along with one or both strains. He noted that each lymph node cell could inhibit one strain or the other but never both. This led to the theory that each antibody-producing cell could produce only one antibody.

Antibody proteins were first isolated by Kabat et al. in 1938 by gel electrophoresis of the protein component of antibody-containing sera before and after mixing with antibody-sequestering antigen. He noticed that the sequestered sample was missing the band at ~150kd, which he dubbed the γ-globulin. In 1961 Edelman published a paper describing how the γ-globulin protein dissociated into two components under reducing condition, a large protein dubbed the heavy chain at ~50kd and a small protein dubbed the light chain at ~25kd. A simple calculation revealed that antibodies were likely to consist of two heavy and two light chains. Further mysteries emerged when it was revealed that the amino acid composition of these chains are variable and that this variation was present at NH2 end of the peptide but not the COOH end.

In 1965 Bennett published a paper describing what was then considered to be a fundamental paradox: How is it possible for one gene to produce an endless variety of protein products with this unique pattern of mutations? Bennet argued that the conserved domain was the product of a single gene while the variable portion was the product of a different set of genes and that the proteins were a product of homology-based binding between these two parts. It was quickly determined that it would be impossible for each unique antibody to be the product of a different gene since the genome did not possess enough physical space to accommodate enough antibody genes to offer the near-limitless level of protection against pathogens observed thus far. In addition, a comparison of Human, Mouse and Rabbit antibodies indicated an evolutionary pattern which would be impossible to replicate via

thousands of independently evolving genes. In 1967 smithies proposed a genetic model whereby there existed a single heavy, kappa, and lambda gene, but for each gene there existed and associated “scrambler” gene which annealed to the variant half of the “master” gene during DNA replication, inducing somatic mutations. Leroy Hood came close to realizing the truth when, in 1970, he proposed a model where the constant and variable regions of each chain where encoded by two separate classes of genes and then combined to form the complete antibody (McKean, Bell, and Potter 1978). M. Weigert noted that variants present in the variable region of Lambda chains were more likely to occur inside three “specificity” regions of the chain. He postulated that many of these mutations were likely to be the result of somatic variants and that they were driven by the antigenic specificity of the antibody and that the true number of variable segment genes was much lower than what had been previously proposed (Weigert et al. 1970).

In 1976 Hozumi et al in the lab of Susumu Tonegawa published a paper describing a series of experiments wherein radio-labeled whole and fragmented kappa chain RNA was hybridized to restriction-digested embryonic and plasmacytoma DNA from mice (Hozumi and Tonegawa 1976). He noticed that whole and 3’ fragments of RNA hybridized to a single band in the plasmacytoma DNA while in embryonic DNA the whole RNA hybridized to two different bands and the 3’ fragments hybridized to the larger of those two bands. He hypothesized that these patterns appear because the kappa chain gene exists as two separate and remote C and V genes in the germline but are joined together through somatic rearrangement

in antibody-producing cells. One year later, direct visualization of kappa mRNA hybridization to myeloma DNA by electron microscopy indicated that, in the rearranged genome, the C and V genes were 1250bp apart (Brack and Tonegawa 1977). Further experiments on the Lambda chain genes revealed the presence of an extra antibody gene segment between the C and V segments which had up to that point remained undetected and which they called the J segment (Brack et al. 1978). Hood et al. identified via sanger sequencing a third segment involved in the rearrangement of heavy chain genes which he called the D segment (Early et al. 1980). In addition, he identified a set of conserved noncoding sequences 3' to the V segment and 5' to the J segment which consist of nearly identical 7 and 10 nucleotide sequences separated by either 11 or 22 nucleotide spacer consisting or random nucleotides. Furthermore, the conserved sequences at the 3' of the V and 5' of the J are nearly inverse complements of one another. Hood correctly hypothesized the presence of a protein which would recognize these sights and initiate recombination.

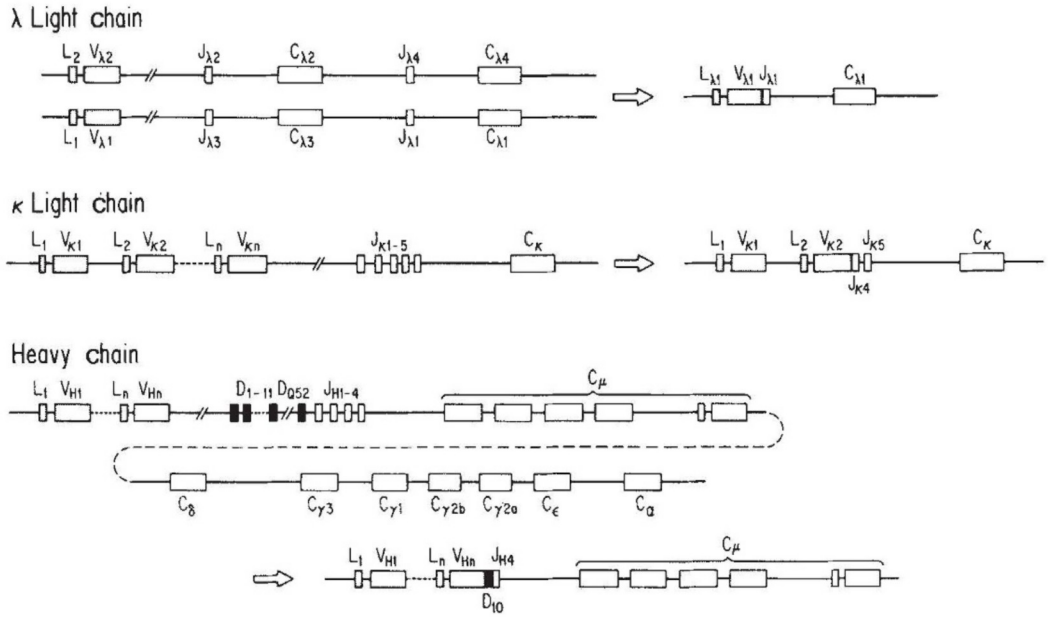

Diagram of Antibody Genetics (Tonegawa 1983)

L. LuzzatI studied the phenomena of multi-class expression whereby B cells express heavy chains which are identical except for the constant region (Pernis, Forni, and Luzzati 1977). He observed patterns in Isotype expression of LPS-treated lymphocytes which suggested that all cells start out with the ability to produce simultaneous IgM and IgD and gradually transition over to producing IgG. The complete picture would be revealed by Kataoka with the allelic deletion model(Honjo and Kataoka 1978). Hybridization experiments revealed that myeloma tumor cells had different copy numbers of different IgG alleles depending on which Isotype they produced. He observed that cells which produced IgG1 had half as many copies of IgG3, and that cells which produced IgG2b had half as many copies of both IgG3 and IgG1. Patterns such as these led him to the conclusion that class-switching

occurs by deletion of the constant region genes between the Variable segment and the new class.

In 1989, the Recombination Activating Gene(RAG-1) was isolated (Schatz, Oettinger, and Baltimore 1989). This was the first gene proven to confer recombination activity at Recombination Signal Sequences(RSS) sites. It was detected by repeated fragmentation and transfection of DNA from a cell line with high recombinase activity into a non-recombining cell line and then plating on media which selected for recombination activity. The transfected DNA was ligated along with a small oligo which the experimenters used as a priming site for sanger sequencing. Further experiments revealed that RAG-1 was evolutionarily conserved to the point where Human RAG-1 cDNA could be used as a probe to detect the RAG-1 genes in mouse, horse, goat, rabbit, and dog as well. One year later RAG-2 would be discovered, a neighboring gene which, in combination with RAG-1, initiate efficient recombination at RSS sites(Oettinger et al. 1990).

By this point it was understood which genes were needed for recombination to occur, however this was not the full story. It was known that nucleotides were incorporated at the sites of recombination which could not have originated from any of the available gene segments. In addition, it was known that the heavy and light chain genes could mutate during cell replication. In 1993 the results of several knockout experiment showed that an absence of Terminal deoxynucleotidyl transferase expression eliminated the incorporation of non-template nucleotides at the CDR3 (Komori et al. 1993). a few years later Activation-induced cytidine

deaminase was shown to be necessary for both somatic hypermutation of antibody genes and class-switch recombination (Muramatsu et al. 2000) and finally, In 1998, a complete copy of the Human Heavy Chain locus was assembled(Matsuda et al. 1998). The ~1.4 Mb locus is notable for being internally repetitive and for containing numerous V segment pseudogenes.

Modern understanding of antibody genetics

In combination these studies laid the groundwork for our current understanding of how the unprecedented diversity of antibody proteins.

In summary:

1.) Each B cell produces an antibody which consists of a pair of unique heavy chains and light chains, each of which can be divided into a variable and constant region.

2.) The Variable region of heavy and light chains consist of 3 Complementary Determining Regions(CDRs) which determine antigen binding and 3 Framing Regions(FRs) which give the antibody structure. At the mRNA level, the variable Variable region occurs at the 5' end of the transcript with the CDRs sandwiched between the FRs. CDR1 and CDR2 are inherited from the V gene segment while the CDR3 is a product of somatic recombination.

3.) During somatic recombination, one V, D (heavy chain only), and J gene segments are randomly selected from a collection of segments in the respective gene locus. These segments are recombined into a variable region. During this process non-template nucleotides are added between the gene segments and nucleotides removed via chewback, generating large amounts of sequence diversity.

4.) Further diversity is generated through somatic hypermutation during affinity maturation and clonal expansion. This process generates a lineage of antibodies which share a common parent but which may possess distinct mutations which affect their ability to bind to antigens. Antibodies with higher affinity to the target antigen will divide faster, driving the production of antibodies with increasing affinity.

5.) The constant region consisted of a single isotope-determining gene segment which could be expressed in either the membrane-bound or secreted form. In addition, the isotype of an antibody could change due to isotype switching. The isotype influences the type of receptors the antibody binds to and the multiplicity of the antibody in its secreted form.

High-throughput sequencing to investigate antibody transcripts

One of the primary technical challenges which faced early antibody geneticists was figuring out the sequence of distinct antibodies. Sanger sequencing was the method of choice for most scientists and this method required two things: 1. A pure DNA template 2. A known priming site on the template. However, any mRNA or cDNA sample derived from a healthy population of B cells will contain contain a mixture of distinct antibody transcripts. Thus, many early studies were done using lymphoma antibodies. Lymphoma are often derived from a single B cell which means that the antibody transcripts derived from the cancer will be relatively pure. Cloning enabled the isolation of individual antibody transcripts, however this method was time-consuming and low-throughput.

Unfortunately, the advent of high-throughput sequencing didn't help address this

challenge. Regular RNAseq protocols would not allow for the identification of distinct heavy and light chains in bulk samples. Because portions of a sequence would be shared between different antibodies while other portions would be unique, it would be impossible to reconstruct antibodies using an Overlap-Layout-Consensus(OLC) or debruijn graph assemblers, which rely on the fact that sufficiently large stretches of the target sequence are unique and can be identified by alignment or k-mer analysis. In fact, the only difference between two independently-generated antibody transcripts may be as small as one nucleotide! Thus, any bulk repertoire sequencing approach would have to generate reads which would both cover the entire variable segment and also be grouped by the original molecule which the sequence was derived from.

In 2011 a method for high-throughput sequencing of antibody repertoires was published (Weinstein et al. 2009). In this study, cDNA amplicons were generated from Zebrafish mRNA by priming off of the second framing region and the constant region in heavy chain transcripts and then sequencing using 2x230 runs on the 454 pyrosequencer. These sequences would cover the entire V segment and a portion of the C segment using two paired reads which could be reconstructed into an antibody transcript by OLC assembly. This provided enough information to identify the isotype and clonality of the heavy chains in a small population of zebrafish. The authors found that the recombination rate for different gene segments varied dramatically and that the rate of V segment usage varied significantly between individuals. That same year a paper was published showing that repertoire sequencing could be used

to detect minimal residual disease in patients with CLL. The authors showed that high-throughput sequencing of heavy chains in patients with CLL could readily identify the presence of cancer-derived antibody clones (Logan et al. 2011). In 2012 repertoire sequencing was applied to the problem of haplotyping the human heavy chain locus (Kidd et al. 2012). The authors did this by taking advantage of the fact the J6 is the most commonly used J segment and that if someone is heterozygous for J6 then heterozygous V and D segments will segregate with one copy of J6 or the other when V(D)J recombination occurs. These patterns can be used to phase gene segments in the heavy chain locus. Soon after it was demonstrated that repertoire sequencing could be used to measure the adaptive immune response through detection of heavy chains conserved across multiple administrations of influenza vaccine (Vollmers et al. 2013). In 2013 repertoire sequencing was used to identify novel HIV neutralizing neutralizing antibodies. Repertoires were generated from HIV+ patients and sequences compared to known HIV neutralizing antibodies using phylogenetic analysis. This method was capable of identifying antibodies with low sequence homology but strong comparable HIV neutralizing ability. Several years later it was shown that repertoire sequencing could be used to detect graft rejection in heart transplant patients undergoing immune suppression and that it was capable of detecting rejection sooner than the cell-free DNA assay which was widely considered to be the gold standard. It was found that graft rejection correlated strongly with the abundance of highly mutated antibody transcripts, especially IgG and IgA sequences (Vollmers et al. 2015).

Even though Immune repertoire sequencing has been used successfully to answer basic and applied questions in immunology, there were still limitations with this technology that my work as a graduate student aimed to address.

Aims

Aim 1: Full Length Heavy Chain Sequencing

Heavy chain transcripts are typically somewhere between 1,500 and 2,000bp long depending on if they are membrane-bound or secreted. Furthermore, about 700bp needs to be sequenced in order to identify the complete variable region as well as enough the the constant region in order to identify isotype. At the time this project was conceived, all published protocols for the sequencing of antibody repertoires involved priming off one of the conserved framing regions in the V-segment, leaving one or more of the CDRs unsequenced (Georgiou et al. 2014). The CDRs define what antigens will bind and with what affinity. If you want to study the antibody or use it in a scientific or healthcare setting, you have to know the sequence of all the CDRs.

To address these limitations, we set out to create a method for sequencing full-length heavy chain transcripts. We accomplished this by using a combination of molecular indexing and random shearing via TN5 to generate a library where entire transcripts can be reconstituted practically error-free.

Aim 2: Heavy and Light Chain Pairing

There are several billion B cells in each person's blood and hundreds of billions of B

cells residing in a person's lymphatic tissue, which means there are billions of unique heavy and light chains which comprise an individual's antibody repertoire. We can infer the amino acid sequences of a person's heavy and light chains by sequencing the mRNA extracted from that person's B cells. However, at the time this project was conceived, it was impossible to know which pairs of heavy and light chains were produced by the same B cell without using single-cell sequencing. In 2015 DeKosky et al. published a method for isolating single B cells in droplets and generating Heavy and Light chain amplicons which could later be sequenced (DeKosky et al. 2015). However, this method resulted in a high dropout rate which was likely due to the stochastic nature of isolating single cells using droplets. We were interested in pairing heavy and light chains for two reasons. From a practical point of view, you need to be able to pair heavy and light chains if you want to identify complete antibodies using repertoire sequencing and without the use of single-cell techniques. Thus, the development of an efficient and robust pairing technique would greatly aid in the expression and analysis of antibodies. In addition, pairing heavy and light chains would allow for the study of chain pairing patterns, biases and interactions.

We proposed that by taking a sample of B cells, dividing it into several smaller samples and sequencing each of these subsamples it is possible to discover heavy and light chain pairs which were present in the original sample. Our method calls for sequencing subsamples containing several thousand B cells and looking for heavy and light chain sequences which occur in multiple subsamples. Because the human body contains millions of B cells, each with their own unique heavy and light chains,

it is highly improbable that any particular heavy and light chain pair would be present in multiple libraries unless they were produced by the same clonal B cell population. This project sought to build upon the work done by Howie et al. in their 2015 paper on pairing alpha and beta T cell receptor subunits using a very similar method, only with heavy and light chains of the antibody (Howie et al. 2015). However, there are several important differences which make this project more difficult. The biggest difference is that the T cell receptors do not undergo somatic hypermutation. Thus, well occupancy can be determined for each alpha and beta subunit by simple sequence identity. Heavy and Light chains must first be clustered into lineages which consist of clonally related sequences. In addition, the unequal expression of heavy and light chains increases the probability of antibody dropout and false discoveries.

By good fortune we developed another technology which proved to be the key to pairing heavy and light chains. Around the same time we were working on a method for sequencing the 5’ ends of transcripts in single cells. At the time, the available technology for single cell RNAseq only supported the sequencing of 3’ ends, and since the 5’ end contains information about the transcription start site having the ability to sequence them would be useful. We developed a simple protocol using poly-A selection and template-switch reverse transcription to generate cDNA molecules with distinct priming sites at the 5’ and 3’ ends, and then followed with tagmentation and amplification to generate libraries which cover the 5’ ends of transcripts. We applied this method to a population of B cells and found that the data could be used to generate paired heavy and light chains. As you may recall, the

variable region of the antibody is present at the 5’ end of the transcript. We found that, with sufficient depth of sequencing, this method provided enough coverage of that region to assemble the variable portion of heavy and light chain transcripts in individual B cells and also estimate their isotype and expression levels. In the end, we were able to pair heavy and light chains as a byproduct of a method we developed for the sequencing of the 5’ ends of transcripts in single cells which we refer to as TN5Prime.

Aim 3: Repertoire Sequencing Using the Oxford Nanopore

Although dedicated antibody sequencing protocols can be used to accurately characterize the repertoire from mRNA, a fundamental problem is that these methods must be used in conjunction with regular RNAseq in order to fully characterize the transcriptome of a population of B cells. However, we demonstrated in the TM5Prime paper that repertoires can be acquired as a byproduct of certain types of mRNA sequencing. We believe that the future of repertoire sequencing and analysis will be as a byproduct of the sequencing of whole transcriptomes. We demonstrate that this is possible by applying the previously published R2C2 method to bulk PBMC RNA. We show that this method can be used to comprehensively measure the transcriptomes of the cells in our population including antibody and T-Cell receptor transcripts.

In the following chapters I will present work I have done towards meeting these research goals. The first two chapters will come in the form of published papers

detailing experiments and their results which meet the technical objectives of the first two aims, and the final chapter will present unpublished results as well as a brief analysis.

Aim 1: Full Length Heavy Chain Sequencing

[THIS SECTION ADAPTED FROM **Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons using Tn5 enabled and Molecular identifier guided Amplicon Assembly**](Cole et al. 2016)

Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier-Guided Amplicon Assembly

Charles Cole[1], Roger Volden[1], Sumedha Dharmadhikari[1], Camille Scelfo-Dalbey[1], Christopher Vollmers[1,*]

Abstract
Antibody repertoire sequencing is a powerful tool to analyze the adaptive immune system. To sequence entire antibody repertoires, amplicons are created from antibody heavy chain (IGH) transcripts and sequenced on a high-throughput sequencer. The field of immune repertoire sequencing is growing rapidly and the protocols utilized are steadily improving, yet, thus far, immune repertoire sequencing protocols have not been able to sequence full-length immune repertoires including the entire IGH Variable region and enough of the IGH Constant region to identify isotype subtypes. Here we present a method that combines Tn5 transposase and

molecular identifiers for the highly accurate sequencing of amplicons longer than 500bp using Illumina short read paired end sequencing. We then apply this method to antibody heavy chain amplicons to sequence the first highly accurate full-length immune repertoire.

Introduction

Antibodies are encoded by Heavy Chain (IGH) and Light Chains (IGκ/λ) loci which undergo somatic recombination during B cell differentiation. In the heavy chain, VDJ recombination creates a highly diverse Complementarity Determining Region 3 (CDR3) when randomly and imperfectly combining one each of ~40 V, ~30 D, and 6 J segments. Heavy chain loci are further modified by somatic hypermutation and class-switch recombination. Somatic hypermutation introduces mutations and indels that can affect the binding characteristics of an antibody. Class-switch recombination changes the antibody isotype by genomic rearrangement of the isotype-determining Constant regions (IgM, IgD, IgG1-4, IgA1-2, IgE). The isotype of an antibody changes the characteristics of an antibody like the ability to bind complement, pass the placenta, or bind certain Fc receptors. Together, VDJ recombination, somatic hypermutation, and class-switch recombination create a virtually unique IGH locus in every mature B cell clone.

Because every B cell clone is unique and can expand and mutate in response to an antigen, analyzing the repertoire of IGH transcripts in a blood sample provides insight into the composition and state of the adaptive immune system. Immune repertoire sequencing has so far been used in both basic and translational research.

In basic research it has been applied to estimate the absolute size of the B cell repertoire in humans, track the effect of aging on the immune system, investigate V,D, and J pairing, and haplotype phase the genomic IGH locus. On the translational side it has been used to track immune response to diseases and vaccines, to track minimal residual disease in leukemia, and determine rejection events following organ transplantation(Boyd et al. 2009; Vollmers et al. 2013; Jiang et al. 2013; Glanville et al. 2011; Arnaout et al. 2011; Meyer et al. 2013; Vollmers et al. 2015). All these studies rely on capturing the diversity of the IGH repertoire but are limited by current sequencing technologies and protocols. Therefore, they either: 1.) utilize long read platforms (454) which allow for the sequencing of the whole IGH variable region plus partial constant region(Glanville et al. 2011; Jiang et al. 2013; Boyd et al. 2009) but are often limited by high cost, lower throughput, and high error rates or 2.) utilize short read sequencers (Illumina HiSeq, Illumina MiSeq, IonTorrent PGM) which allow for higher throughput at lower cost and error-rate but are limited by their short read length to sequence only part of the IGH Variable Region including the CDR3 (Vollmers et al. 2013; Meyer et al. 2013). To provide complete information of an IGH repertoire an ideal full-length IGH transcript amplicon would be ~530bp in length, starting at the Leader exon and ending at 100bp into the Constant Region. Starting in the Leader exon, which does not encode for the final antibody protein, would ensure no bases in the Variable region are masked by primers, which would ensure the accurate identification of all V segment alleles. Ending 100bp into the Constant Region would provide enough sequencing information to distinguish all Isotype and subtypes, which is essential for allergy research.

While an Illumina MiSeq 2x300 sequencing run theoretically allows for the sequencing of this ideal 530bp IGH amplicon, in practice declining base quality doesn't allow for the sequencing of an amplicon longer than 450bp. Recently, several groups have employed approaches utilizing molecular identifiers (UIDs) to improve sequencing accuracy which is essential to differentiate somatic hypermutation from PCR and sequencing errors (Vollmers et al. 2013; Shugay et al. 2014; He et al. 2014). Further, several groups have developed protocols utilizing short read sequencers to sequence individual molecules exceeding the current raw read length of these sequencers. These protocols rely on inefficient steps in library preparation including Biotin pulldowns and intra molecular circulations (Hong et al. 2014; Hiatt et al. 2010; Wu et al. 2014; Rossano et al. 2009; Lundin et al. 2013). To overcome read length, accuracy, and library preparation limitations, we developed Tn5 enabled Molecular Identifier guided Amplicon sequencing (TMIseq). TMIseq is based on a simple library preparation protocol utilizing molecular barcoding of individual molecules and Tn5 tagmentation (Picelli, Björklund, et al. 2014) enabling the highly accurate and cost effective sequencing of molecules exceeding Illumina read length (Fig. 1).

Results

Overview of TMIseq

To assemble RNA molecules that exceed the sequencing length but not the cluster generation length of Illumina sequencers, TMIseq utilizes molecular identifiers and

the unique characterisitics of the Tn5 enzyme. We reverse transcribe RNA molecules into cDNA and then generate 2nd and 3rd strand copies of cDNA in a two cycle amplification reaction using two primer pools. The primer pools we used for to assemble IGH RNA molecules were: 1.) The V_Leader pool containing primers specific to the Leader exons of all V segments 2.) The C_long pool containing primers that bind 100bp into the Constant regions of all Isotypes (C_long). All primers in these pools feature modified 5' ends to generate a single 3rd strand cDNA copy of each IGH RNA molecule tagged with 18bp random molecular identifiers and partial Nextera (Illumina) sequences on both ends (Nextera_A for V_Leader, Nextera_B for C_long) (Fig. 1). We then amplify these uniquely tagged cDNAs using primers specific to the partial Nextera sequences that preserve the molecular identifiers and add dual-indexed complete the Nextera sequences. This dual-indexed ~530bp amplicon library is at this point Illumina-sequencing-ready. We then split the library into three aliquots. The first and second aliquots (*Tn5_A* and *Tn5_B*) are tagmented with Tn5 enzyme loaded only with partial Nextera_A (*Tn5_A*) or Nextera_B (*Tn5_B*) oligos and PCR amplified to complete the Nextera_A (*Tn5_A*) or Nextera_B (*Tn5_B*) sequences, respectively (Fig. 1). The third aliquot (*Uncut*) is left unchanged and sequenced alongside the *Tn5_A* and *Tn5_B* libraries (Fig. 1). Illumina chemistry only sequences molecules with both complete Nextera_A and Nextera_B sequences at their ends (Fig. 1). Therefore, *Tn5_A* and *Tn5_B libraries* exclusively produce raw read pairs in which one read is anchored by the V_Leader (*Tn5_A*) or C_long (*Tn5_B*) primers and contains one of the molecular identifiers associated with the original template molecule, whereas the other read is primed

from the Nextera sequence introduced at a random location into the amplicon by Tn5. Finally, the *Uncut* library exclusively produces raw read pairs in which both reads are anchored by V_Leader or C_long primers and contain both molecular identifiers associated with the original template molecule (Fig. 1). For analysis, after quality trimming and filtering, *Uncut, Tn5_A* and *Tn5_B* read pairs containing highly similar molecular identifiers (Fig. S1) in their anchored reads are combined into IGH molecule groups. IGH molecules are then assembled from each group using AMPssembler, a custom k-mer based amplicon assembler that takes advantage of the known properties of the TMIseq protocol. Namely, the ends of the assembled sequences are defined by the anchored reads and there is only a single sequence to be assembled per IGH molecule group.

Figure 1

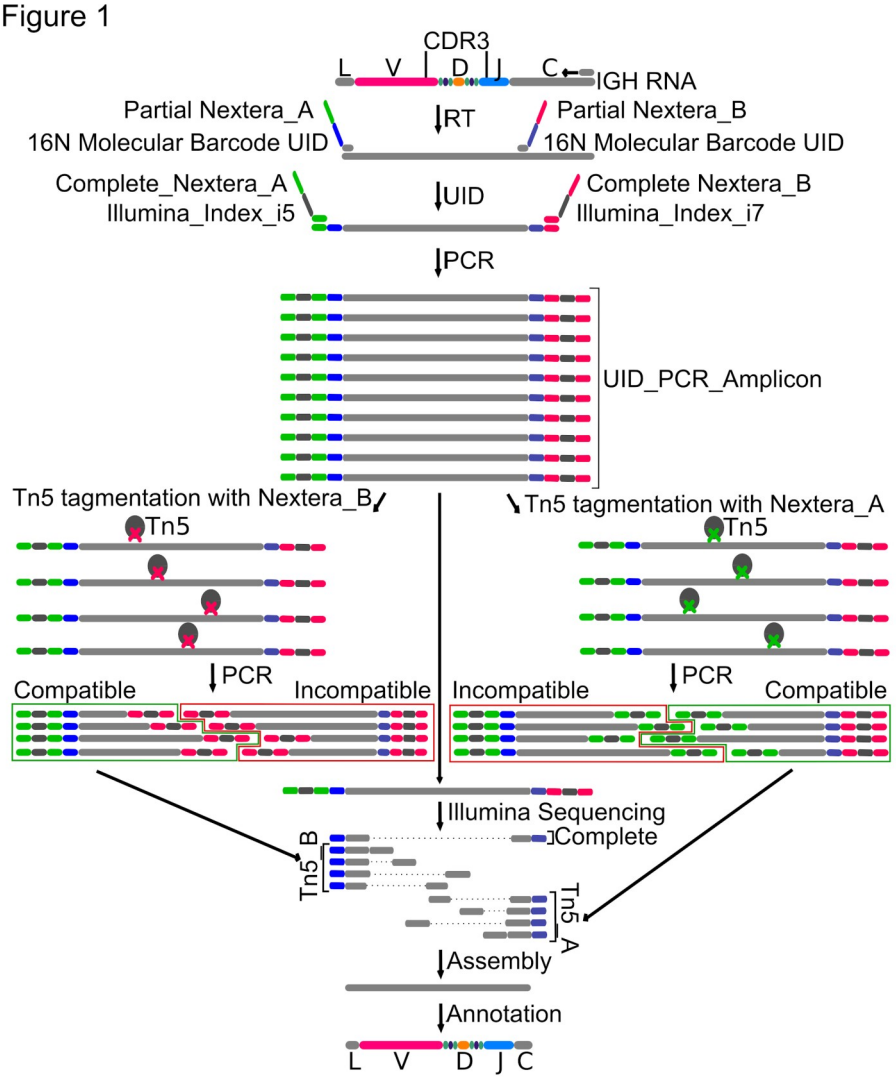

Figure 1: Schematic TMIseq Library Preparation and Data Analysis.
IGH RNA is reverse transcribed and second and third strand cDNA is generated using 5prime modified primers. After PCR amplification the amplicons are tagmented using custom loaded Tn5 enzymes. 3 libraries per sample are sequenced and the resulting reads are grouped using molecular identifiers and assembled with a custom algorithm (AMPssembler).

Application of TMIseq to the analysis of IGH transcript amplicons

To test the TMIseq protocol and data analysis, we created TMIseq libraries from two

individuals (I1 and I2) from samples of PBMCs (Peripheral Blood Mononuclear Cells)

which contain B cells. For I1, we generated TMIseq *Uncut*, *Tn5_A*, and *Tn5_B* libraries for one sample (I1 L1), sequenced those libraries on a MiSeq 2x300 run and truncated the resulting reads to 150bp to model the shorter read length. The MiSeq run generated 125,200 raw reads for the I1 libraries, which yielded 120,104 quality trimmed reads. The trimmed read pairs were assembled by AMPssembler into 2779 IGH molecules. For I2, we generated TMIseq *Uncut*, *Tn5_A*, and *Tn5_B* libraries for 8 samples (I2 L1-L8) and sequenced those libraries on a HiSeq3000 2x150 run. The HiSeq run generated 15,587,484 raw reads pairs across the 8 I2 samples, which yielded 10,577,945 quality trimmed read pairs. These trimmed read pairs were assembled by AMPssembler into 115,108 IGH molecules (11,075-16,985 per library).

TMIseq coverage requirements

We used I2 L1 data to determine the coverage requirements to assemble IGH molecules and enable future optimization of raw read depth. Successful assembly was strongly dependent on read coverage, which itself was strongly positively correlated between the *Uncut* and *Tn5_*A/*Tn5_B* libraries (Fig. 2A, B). TMIseq assembly success increased from 15% for IGH molecules covered by only one *Uncut* read pair to 60-70% for reads covered by 5 or more Uncut read pairs (Fig. 2C), with the assembly success of individual molecules being highly dependent on *Tn5_A* and *Tn5_B* coverage, reaching over 90% for IGH molecules covered by more than 40 combined *Tn5_A* and *Tn5_B* read pairs (Fig. 2C).

Next, we performed rarefaction analysis to determine the ideal coverage levels

required for effective assembly. While subsampling of the *Tn5_A* and *Tn5_B* raw reads had a strong impact on the number of IGH molecules that were successfully assembled, subsampling of the Uncut raw reads had only minimal effect until the number of raw reads fell below 2-5 times the maximum number of assembled IGH molecules (Fig. 2D). A good trade-off between assembled IGH molecules and raw read coverage therefore appears to be 5 *Uncut* raw reads and 30-40 raw reads each for *Tn5_A* and *Tn5_B* for every high abundance IGH molecule in the *Uncut* library. This compares highly favorably to other approaches that enable the sequencing of molecules exceeding the Illumina read length limit (Hong et al. 2014). Further, raw read requirements are likely to be lower if using a HiSeq2500, as the HiSeq3000 has a preference for short molecules, which resulted in ~40% of *Tn5_A* and *Tn5_B* reads to be discarded in a quality filtering step because they were too short or contained adapter sequences.

Figure 2

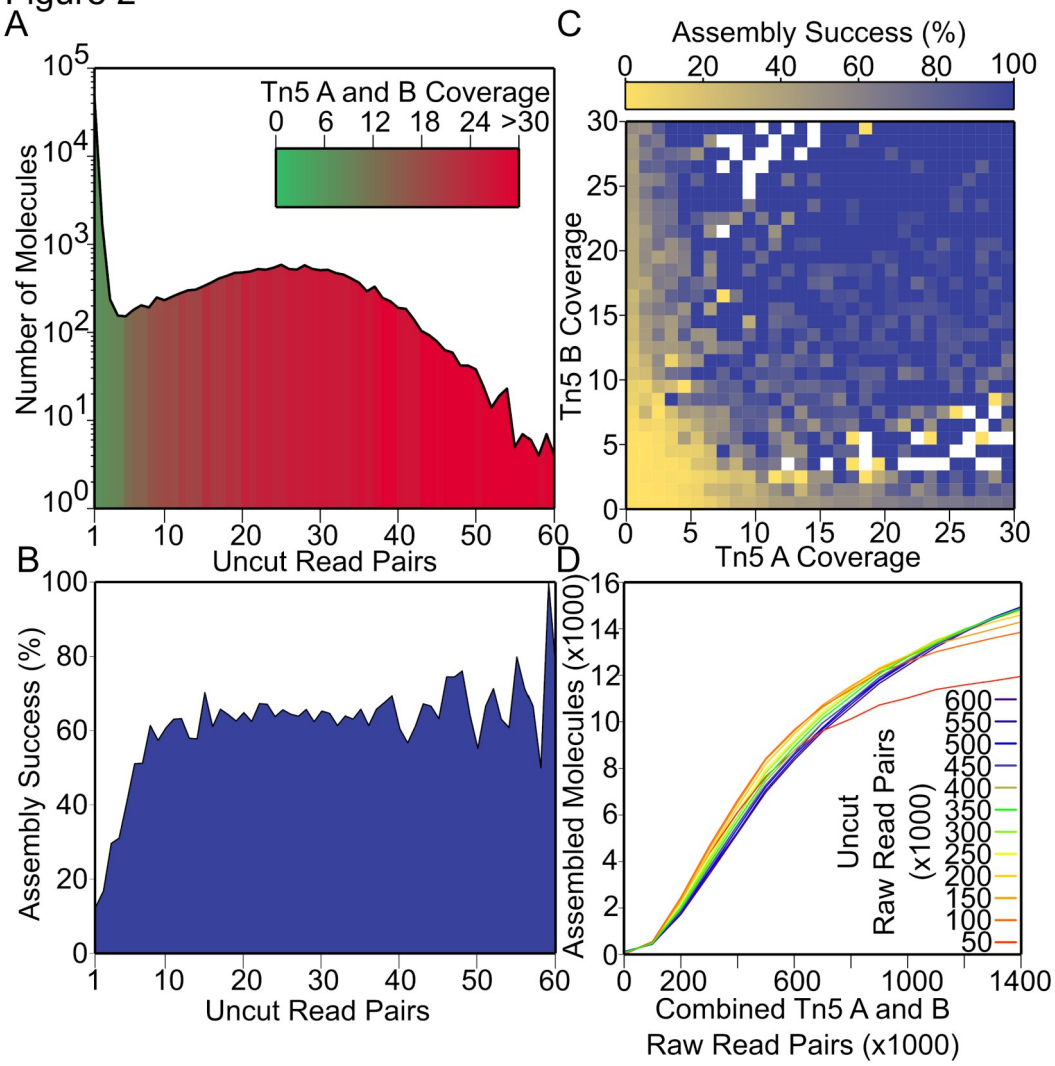

Figure 2: TMIseq Subassembly Coverage Requirements.
A) Read pair coverage for IGH molecules in the I2 L1 Uncut library is shown as a histogram. Average Combined Tn5_A and Tn5_B read coverage at increasing Uncut raw read coverage levels is shown as a color gradient.
B) Average assembly success at increasing I2 L1 Uncut read coverage levels is shown.
C) Heatmap showing the correlation of assembly success and read coverage in I2 L1. Average Success percentage for Tn5_A and Tn5_B coverage combinations is shown.
D) Number of I2 L1 IGH molecules successfully assembled from increasing numbers of subsampled Uncut raw read pairs (line colors) and combined Tn5_A and Tn5_B raw read pairs is plotted.

TMIseq data quality

To assess TMIseq data quality and characteristics we analyzed IGH molecules assembled from the I1 L1 library. The average length of the assembled IGH molecules was 530bp (Fig. 3A) and trimmed Tn5_A and Tn5_B reads aligned to the assembled molecules in the pattern expected based on the library prep protocols (Fig. 3B). Of the 2779 assembled IGH molecules, 98% were identified as heavy chain transcript and annotated by IgBlast (Ye et al. 2013). We then compared these annotated IGH molecules to standard molecular-identifier based immune repertoire control data (I1 Control) derived from a biological replicate and produced using a shorter 400bp amplicon and a 2x300 run on a MiSeq (Fig. 4A).

To assess base-exchange errors we took advantage of IgD sequences which are thought to be expressed almost exclusively by naïve B cells. The vast majority of sequenced IgD sequences should therefore be not mutated. Indeed, we found that most IgD sequences were not mutated: 95.16% of IgD sequences in the I1 L1 TMIseq library and 93.6% of IgD sequences in the I1 Control library showed >99% identity to reference). Most importantly, the percentage of mutated IgD sequences was comparable between TMIseq and error-corrected control libraries

Next, we tried to assess the rates of artificial insertion and deletions of the TMIseq protocol, which, as it relies on computational assembly of sequences, might be prone to generate the kinds of errors. First, we analyzed the observed CDR3 length and potential frame-shifts in the variable region. Lengths of the CDR3s, which is the result of the random recombination of V, D, and J segments and the addition of

quasi-random P and N nucleotides, are expected to occur in steps of three to maintain the reading frame of the antibody heavy chain transcript. Second, we analyzed indels occurring in the rest of the Variable region. Indels in the variable region should occur in multiples of three to result in the addition or loss of whole amino acids, while maintaining the reading frame of the transcripts. We found that the rates of out-of-frame CDR3 (Fig. 4B, C) and frame-shift events in the rest of Variable region (Fig. 4D) were very similar between I1 L1 TMIseq and I1 Control libraries. Together, this confirmed that the rate of errors generated by the TMIseq is equivalent to the very low rates of the error-corrected control protocol (Vollmers et al. 2013).

Figure 3

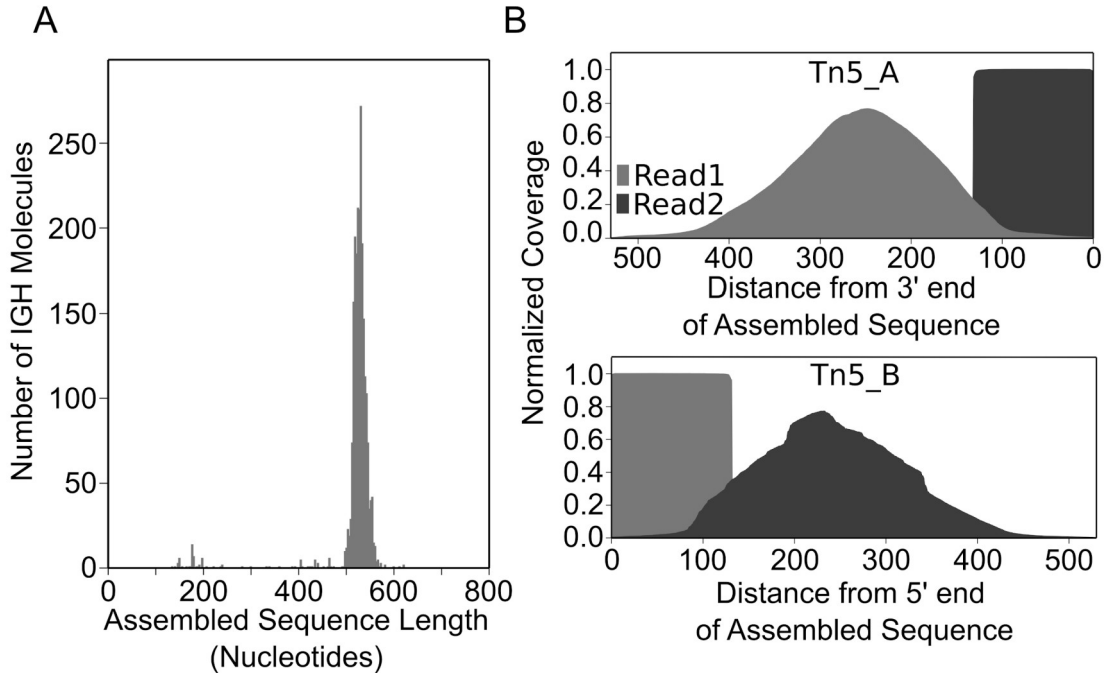

Figure 3: TMIseq Assembles 530bp IGH Molecules

A) Length distribution of I1 L1 IGH molecules assembled using TMIseq.
B) Trimmed Tn5_A and Tn5_B reads are mapped to assembled IGH molecules using BLAST. Mapped read coverage across IGH transcripts is shown as histograms.

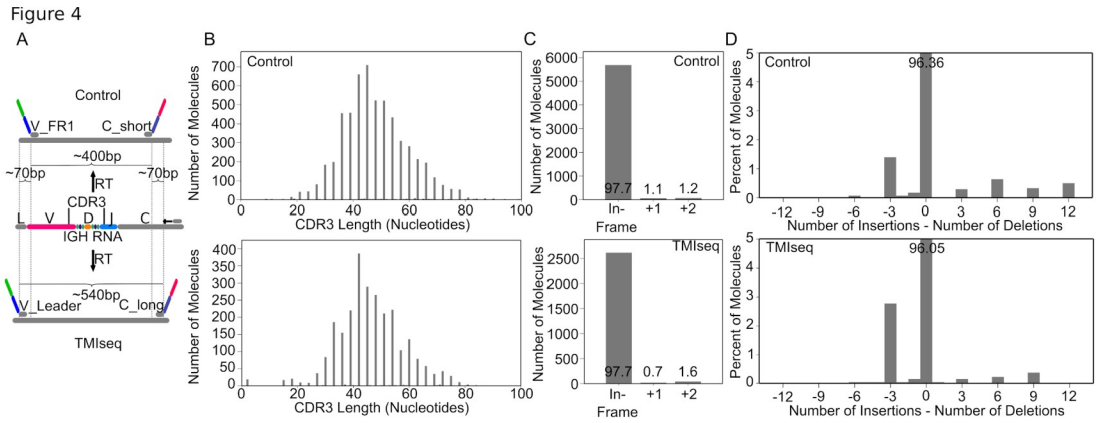

Figure 4: TMIseq Mutations Data Equivalent to Control Libraries
A) Schematic of Primer positioning for second and third strand synthesis in TMIseq and control Libraries. B-D) Control and TMIseq sequences are compared for CDR3 length distribution (B), CDR3 translation frame (C), shift in frame produced by indels (D).

Variable and Constant region coverage by IGH amplicon and TMIseq

The increased sequencing length made possible by TMIseq enabled us to create a longer amplicon by priming in the Leader exon and 100bp into the Constant region. Priming in the Leader exon, which is not included in the final antibody protein allowed us to read every base of the Variable region without it possibly modified by a primer. This enabled us to uniquely identify all V segment alleles. In contrast to the I1 Control library, the I1 L1 TMIseq library was able to identify the V segment allele IGHV3-23*04 that differs from the more common IGHV3-23*01 allele by a single base in the first 20bp of the segment (Fig. 5A). Additionally, priming in the Leader exon enables us to identify mutation hot spots in the entire Variable region, including

potential hot spots in the first 20 bases of the IGHV3 V segments family (Figure 5B). On the other end of the amplicon, priming 100bp into the Constant regions creates an amplicon that contains enough distinct base positions to not only distinguish isotypes like IgM and IgG, but isotype subtypes like IgG1 and IgG3. Indeed, in contrast to the I1 Control Library, the I1 L1 TMIseq library differentiates isotype subtypes including IgG1, IgG2, and IgG3 as well as IgA1 and IgA2 (Fig. 5C). While IgG4 and IgE which are essential for allergy research were only detected at very low levels in the data, this is likely due to the low sequencing depth and their naturally low levels in a mix of IGH transcripts.

Figure 5

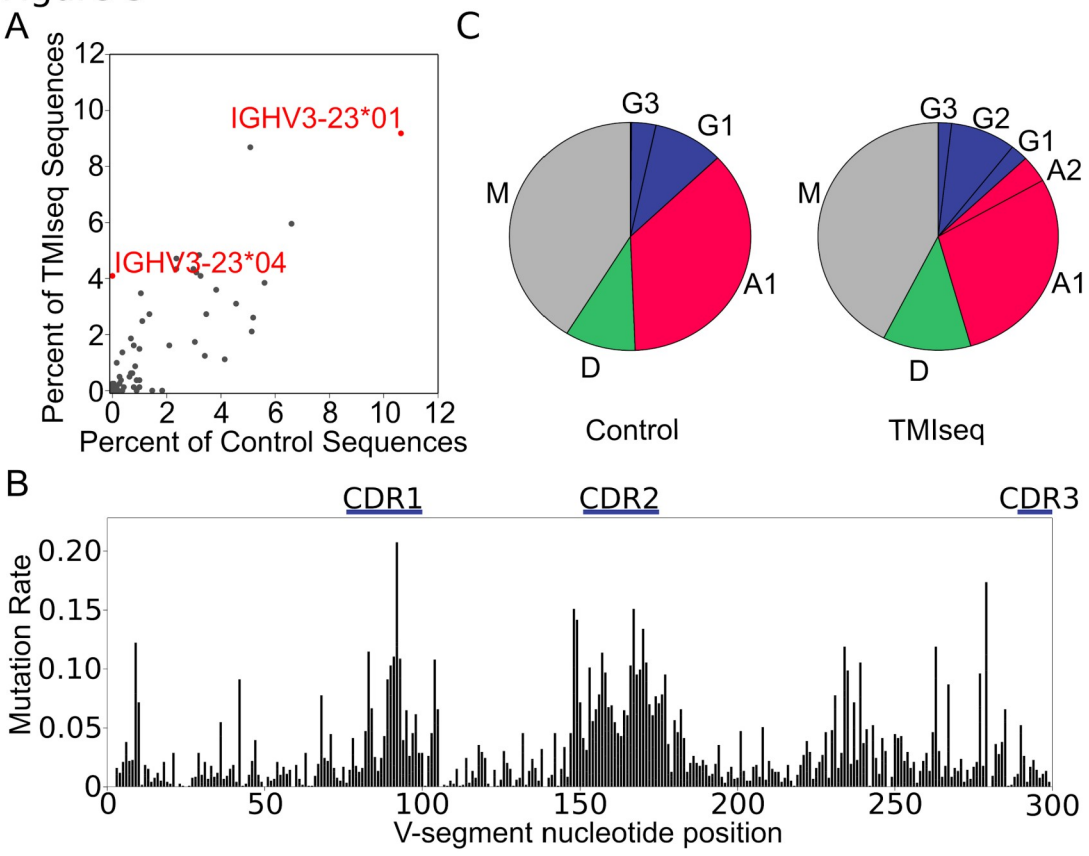

Figure 5: TMIseq can differentiate all V segments and Isotypes
A-B) Control and TMIseq sequences are compared for perfectly matched V segment usage in a scatter plot (A), and Isotype usage in a pie chart (B). C) Mutation rate across the entire V segment averaged across all IGH molecules using V segments of the IGHV3 family.

Finally, to test the data for obvious recurring assembly artifacts and contaminations, we compared IGH molecules derived from all I1 and I2 libraries. Similar to what we had previously shown for standard immune repertoire data (Vollmers et al. 2013), IGH molecule lineages derived from I1 and I2 samples were shared at high levels between the samples of an individual, yet only at very low levels between individuals (Fig. 6), which confirmed the absence of rampant cross contamination and assembly artifacts.

Figure 6

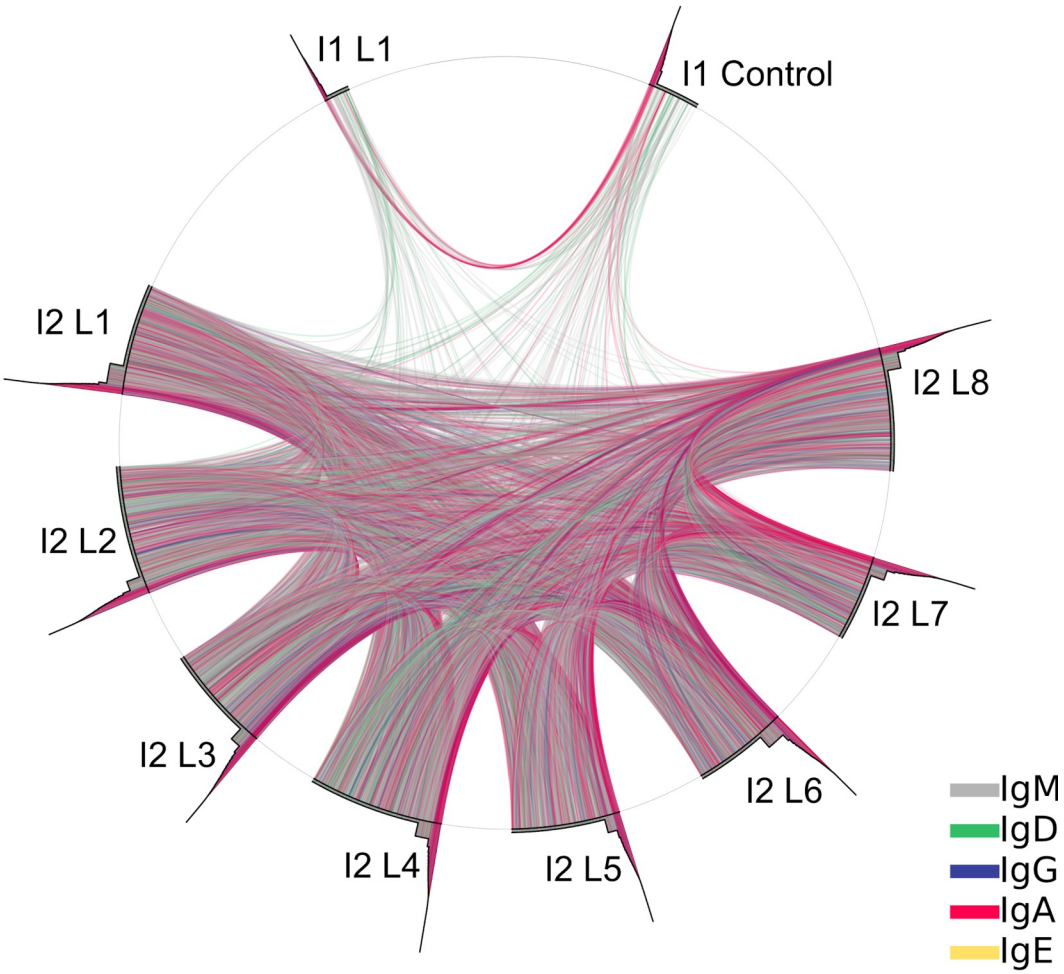

Figure 6: TMIseq Data Identifies Clonal IGH Lineages
Visualization of IGH molecule lineages shared between samples. IGH molecule lineages of each sample are plotted on the circumference of the circle, with the area representing abundance of the respective lineages (logarithmic) and the color representing isotype. IGH molecule lineages present in two time points are connected with lines colored according to their isotype.

Discussion

In this study, we show that the TMIseq protocol we developed enables the sequencing of amplicons that exceed Illumina read length but not cluster generation length. With the current state of technology, this includes amplicons 450bp-800bp in length. We applied TMIseq to IGH amplicons creating an immune repertoire sequencing protocol that is unprecedented in its combination of sequencing accuracy and coverage of Variable and Constant regions. Indeed, there currently exists no other protocol to accomplish this combination. Coverage of the complete variable region will make it possible to query all possible mutations in an antibody and adapt antibody repertoire sequencing for the pairing of heavy and light chains as was recently shown for T cells (Howie et al. 2015). Further, we have shown here that complete variable region coverage, paired with error-correction, enables the identification of highly similar but distinct V segment alleles. This distinction will be essential for the inferred haplotype phasing of the IG loci which was shown previously but hampered by the lower throughput and accuracy of the 454 sequencer (Kidd et al. 2012). Finally, in addition to the coverage of the complete variable region, the increase coverage of the Constant region enables the identification of isotype subtypes for each molecule, which is essential for the study of class-switching and allergies (Looney et al. 2016). The protocol we present is straightforward and cheap to implement by allowing for pooling strategies to minimize the use of Tn5 enzyme. In our hands the complete protocol, from RNA to sequencing libraries, can be completed in a single day. The sequencing cost per TMIseq sample is lower than sequencing a shorter amplicon, which would omit either Variable or Constant region

coverage, on a MiSeq 2x300 run using molecular identifiers for error-correction. While we applied the TMIseq protocol to IGH amplicons, it will be easy to adapt the protocol to any other amplicon. There are several amplicon based applications for which this might be beneficial. These applications include among others the sequencing of 16S RNA and cancer amplicon panels. Taken together, data quality and coverage requirements shown here prove that the TMIseq protocol is capable of sequencing full-length immune repertoires, or any other amplicon between 450bp and 800bp, highly accurately and at high throughput.

Methods

PBMC extraction and RNA purification

All experiments were approved by the Internal Review Board at UC Santa Cruz and Stanford University. For sample I1 whole Blood samples were collected from a healthy human adult volunteer by the UCSC Student Health Center. For sample I2, buffy coat were provided completely de-identified by the Stanford Blood Center. I1 and I2 samples were processed by Ficoll-Gradient (GE-Health) to extract PBMCs. PBMC were lysed directly in RLT buffer and frozen at -80°C until RNA was extracted. RNA was extracted from 400,000 cells each using the RNeasy Mini Kit (Qiagen). Resulting RNA concentrations ranged from 20-50ng/ul.

TMIseq Library Preparation

10ul of RNA was used for Superscript II (Thermo) cDNA first strand synthesis using a primer pool specific to all exons specific to the secreted isoform of all IGH isotypes (IgM, IgD, IgG1-4, IgA1-2, IgE). In a 2 cycle PCR reaction 2nd and 3rd cDNA strands were synthesized using Phusion polymerase (Thermo) and 2 modified primer pools complementary to the beginning of the V-Leader exons and about 100bp into CH1

exons of all IGH isotypes and containing molecular identifiers and partial Nextera Sequences. cDNA was purified and size selected twice with SPRI beads using a 0.7:1 (Beads:Sample) ratio corresponding to a cutoff discarding DNA shorter than 300bp. In a 30 cycle PCR reaction 3rd cDNA strands were amplified using a pair of primers containing complete Nextera sequences as well as Illumina i5 and i7 indexes to index each individual sample. Samples with unique i5 and i7 indexes (i.e. each sample can be uniquely distinguished by either i5 or i7 index, e.g. Sample 1: i5_1, i7_1; Sample 2: i5_2, i7_2; etc...) are pooled and split into three aliquots. To create Tn5_A libraries, aliquot 1 is tagmented using Tn5 enzyme(Picelli et al. 2014) loaded with Nextera_A adapter and PCR amplified using a universal Nextera_B primer and a Nextera_A primer with a Illumina Index not yet present in the library pool and purified and size selected for fragments larger than 380bp using 2% EX Gels (Life). To create Tn5_B libraries, aliquot 2 is tagmented using Tn5 enzyme(Picelli et al. 2014) loaded with Nextera_B adapter and PCR amplified using a universal Nextera_A primer and a Nextera_B primer with a Illumina Index not yet present in the library pool and purified and size selected for fragments larger than 380bp using 2% EX Gels (Life). Uncut (aliquot 3), Tn5_A, and Tn5_B libraries were pooled and sequenced according to standard Illumina protocols on an Illumina MiSeq 2x300 run or HiSeq3000 2x150 run.

Control Library Preparation

Control libraries were generated as TMIseq libraries with the exceptions to the primer pools used for 2nd and 3rd strand cDNA synthesis. The FR1 specific primer pool was designed to bind 1-10bp into the FR1 region, while the C specific primer pool

was designed to bind 20bp into the CH1 exons of all IGH isotypes. The resulting library with an insert size of ~400bp is sequenced on an Illumina MiSeq 2x300bp run.

Raw data processing data assembly

Raw reads in fastq format are trimmed using trimmomatic (Bolger et al. 2014), discarding reads pairs containing adapters. For libraries sequenced on the MiSeq 2x300, reads were also cropped to 150bp. TMIseq data was further processed according to the following pipeline: First, molecular identifiers are extracted from the trimmed fastq files. For Uncut libraries the first 18 bases of read 1 represent molecular identifier 1 and the first 18 bases of read 2 represent molecular identifier 2. For Tn5_A libraries the first 18 bases of read 2 represent molecular identifier 2. For Tn5_B libraries the first 18 bases of read 1 represent molecular identifier 1. Second, reads of the Uncut library are grouped into molecular groups if their combined molecular identifiers differed by less than 5 mismatches. Third, reads with highly similar (less than 2 mismatches) molecular identifier 1 (Tn5_B) or molecular identifier 2 (Tn5_A) to the Uncut molecular groups are added into these molecular groups. Third, the AMPssembler algorithm assembles IGH transcripts from each molecular group. The program sorts the raw reads into three categories: 1.) reads derived from the 5' "end" of the amplicon. 2.) reads derived from the 3' "end", and 3.) reads which are unachored and likely to come from some place in the middle of the molecule. Then the program creates a high-quality consensus of the ends of the amplicon using a combination of quality and abundance of each nucleotide at each position. Finally, the program reduces all of the reads into k-mers and extends one of the ends until the program reaches the other end of the amplicon or the program runs out of

extensions. It then reports the completed molecule in fastq format. When multiple extensions are possible, the program always selects the extension which results in the highest-quality base being incorporated into the extension. Control data was processed as previously described (Vollmers et al. 2013). To analyze molecule coverage distribution in Fig. 2B we aligned the raw reads of each molecule group to the assembled molecule using BLAST (Altschul et al. 1990). Data was then converted to fasta format and annotated using IgBLAST (Ye et al. 2013) with germline data retrieved from IMGT (Lefranc et al. 2004). For Fig. 2C IGH molecules were grouped into lineages across all samples analyzed using a single linkage clustering approach and a 90% CDR3 similarity cut-off. For Fig. 1E reads were subsampled to the approximate target levels from the unprocessed fastq file pairs. The resulting subsampled files were then analyzed by the complete analysis pipeline. Further downstream analysis and visualization was done using Python/Matplotlib (Hunter 2007).

Data Access

The AMPssembler script used in the analysis of the data is available at Github at https://github.com/chkcole/AMPssembler. All other scripts are available upon request.

Raw data was uploaded to the SRA under Bioproject ID PRJNA291102 (I1 data are identified by anonymized ID SHC1-3-1, I2 data are identified by anonymized ID BB7)

Aim 2: Heavy and Light Chain Pairing

[THIS SECTION ADAPTED FROM **Tn5Prime, a Tn5 based 5 capture method for single cell RNA-seq**](Cole et al. 2018)

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]

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

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

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 Smartseq2 RNA-seq protocol, a powerful tool in deciphering the complexity of single

cell heterogeneity(Picelli, Faridani, et al. 2014; Treutlein 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)(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 (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. Therefore, none of the current 5' end specific protocols are capable of efficiently and cost-effectively processing hundreds to thousands of single cells necessary to understand heterogeneity within complex mixtures of cells present in, for example, the adaptive immune system or cancer.

Furthermore, new droplet based high-throughput single cell RNAseq approaches like Drop-Seq and 10X Genomics Chromium platform can process thousands of cells but can only analyze the 3'end of transcripts due to integrating a sequencing priming site into 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.

To overcome this lack of high-throughput single cell 5’ capture methods, we chose to modify the Smartseq2 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 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 (Table S1, Fig. 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 contain 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. 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. 1B,C).

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 the standard Smartseq2 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 a Smartseq2 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 standard Smartseq2 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.

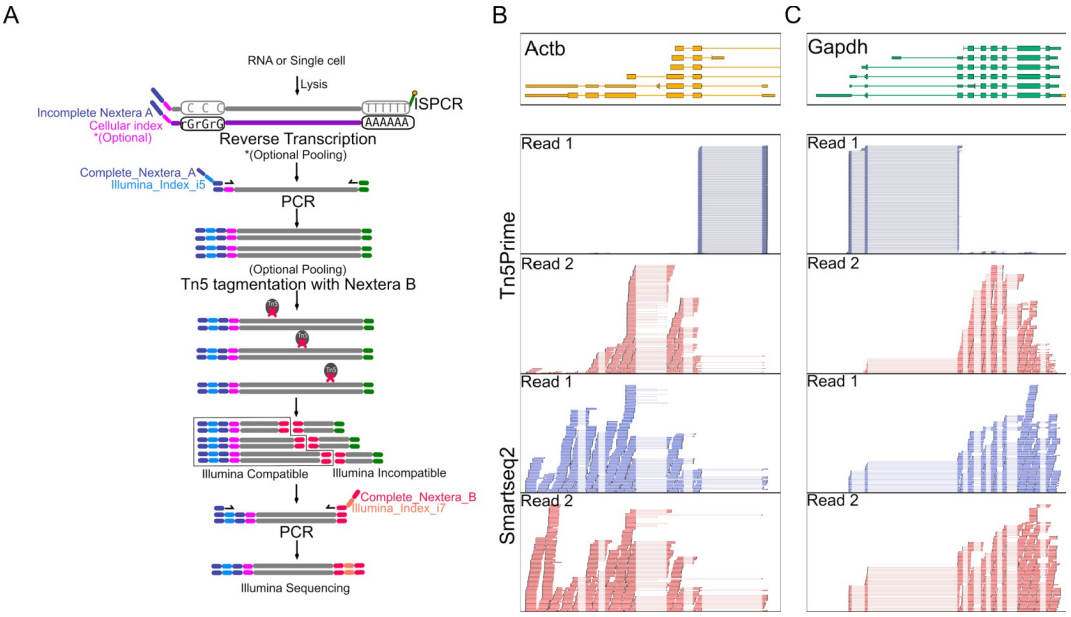

Fig. 1 Tn5Prime Library construction and 5' capture
A.) Schematic of the Tn5Prime library construction. No enrichment steps are required to generate a library that captures the 5' end of transcripts. B.) Read alignment plots comparing 5' end capture by Tn5Prime to random fragmentation by Smartseq2 using 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.)

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 Smartseq2 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. 1B). Next, systematic analysis was based on mapping the start of read 1 to identify putative Transcription 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). 89.7% of the 17853 peaks fell within TSSs (0-25 bp upstream) with the vast majority of them falling near promoter regions (26bp-1000bp upstream) or 5'UTRs (Fig. 2A). Next, we subsampled the CAGE data to levels similar to the Tn5Prime data and called peaks in the same manner. We found that 14107 of 17853 Tn5Prime peaks (73%) fell within 25bp to the nearest of 27526 CAGE peaks, indicating high concordance between the two approaches (Fig. 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. 2B). Ultimately, this suggests that our Tn5Prime protocol is equivalent to the gold standard CAGE technique in targeting transcription start sites.

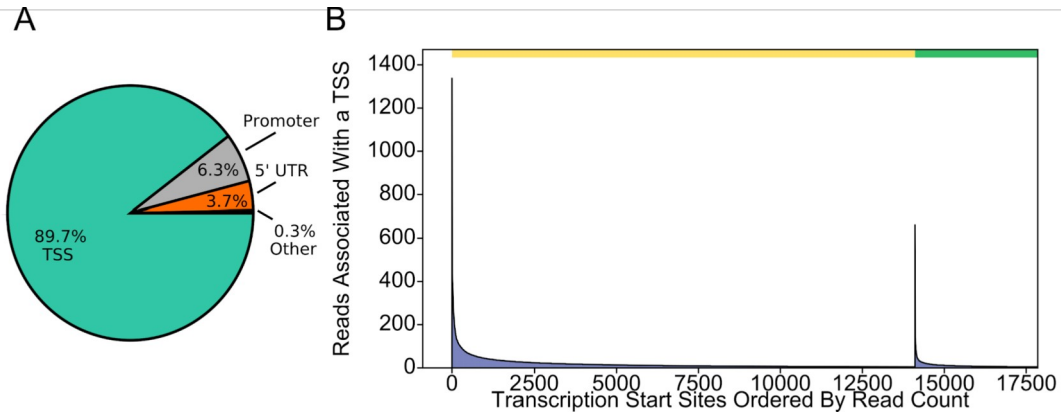

Fig. 2 Tn5Prime peaks are highly concordant with GENCODE annotation and CAGE peaks

A) Tn5Prime peaks identified in lymphoblast cell line GM12878 using 50 ng of input RNA were matched to features determined by Gencode annotation. Features are shown as a pie chart. B) Tn5Prime peaks generated from GM12878 were matched to CAGE peaks also generated from GM12878. The green bar on top indicates the peaks within 25 bp and the yellow 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, Smartseq2 data generated by our lab, as well as CAGE and RNA-seq data produced by the ENCODE project (Fig. 3). We used the Tn5Prime data mentioned in the previous section and generated the Smartseq2 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 Smartseq2 with a Pearson r of 0.87 (95% C.I. 0.86-0.87). Tn5Prime and Smartseq2 data were comparable with ENCODE RNA-seq and CAGE data (Fig. 3), indicating that the Tn5Prime protocol is equivalent to the conventional Smartseq2 method in measuring transcript abundance. Together, these data show that Tn5Prime can accurately identify transcription start sites and quantitatively measure transcript abundance.

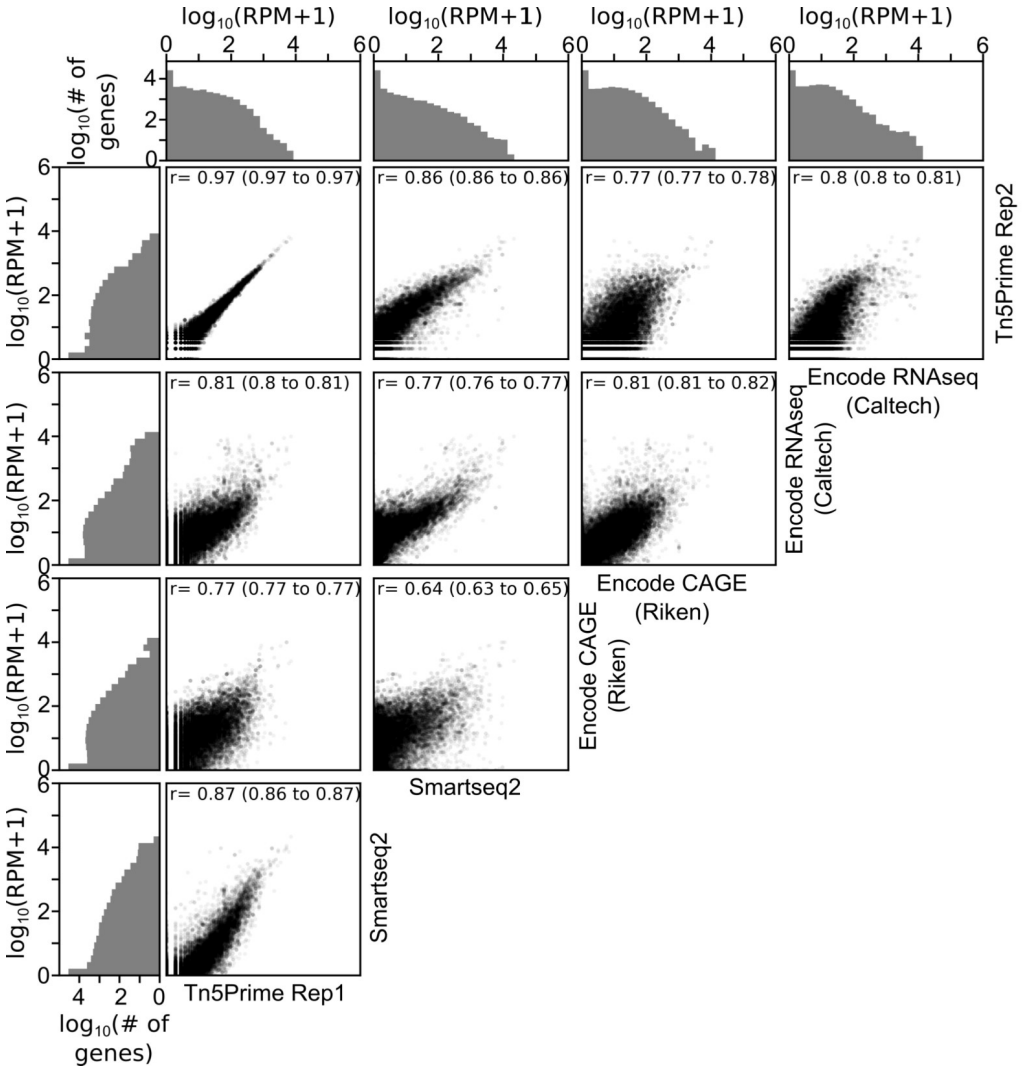

Fig 3. Tn5Prime quantifies transcriptomes accurately and reproducibly.
Pairwise correlations of transcript levels between Tn5Prime, Smartseq2, ENCODE CAGE and ENCODE RNAseq experiments using GM12878 cell line are shown as scatter plots. A total of 50 ng of input RNA was used. Each transcript is shown 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 transcription start site localization in single B cells.

As the Tn5Prime protocol is based on the same cDNA amplification strategy as the Smartseq2 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 previously published data on single B cell RNAseq (Zheng et al. 2017; Jaitin et al. 2014; Gierahn 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 (Fig. S1). These data indicate that we can efficiently 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 (Fig. S2). 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 is defined by high levels of surface markers CD27 and CD38, but separation from memory B cells

which also express these markers, albeit at lower levels, can be challenging. Therefore, analyzing these cell types at the single cell level should help further delineate these populations.

We developed a multiplex strategy by inserting cellular indexes into the template switch oligo which allows the user to pool samples 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.

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 STAR with a median of 74.59% percent of cell-assigned reads being uniquely assigned to an annotated gene. Each cell expressed a median of 534 genes. Of the 58234 annotated genes in GENCODE, 5414 genes had at least one read per cell on average. 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 and Encode CAGE data. 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 (Fig. 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 protocol (Fig. 4C,D).

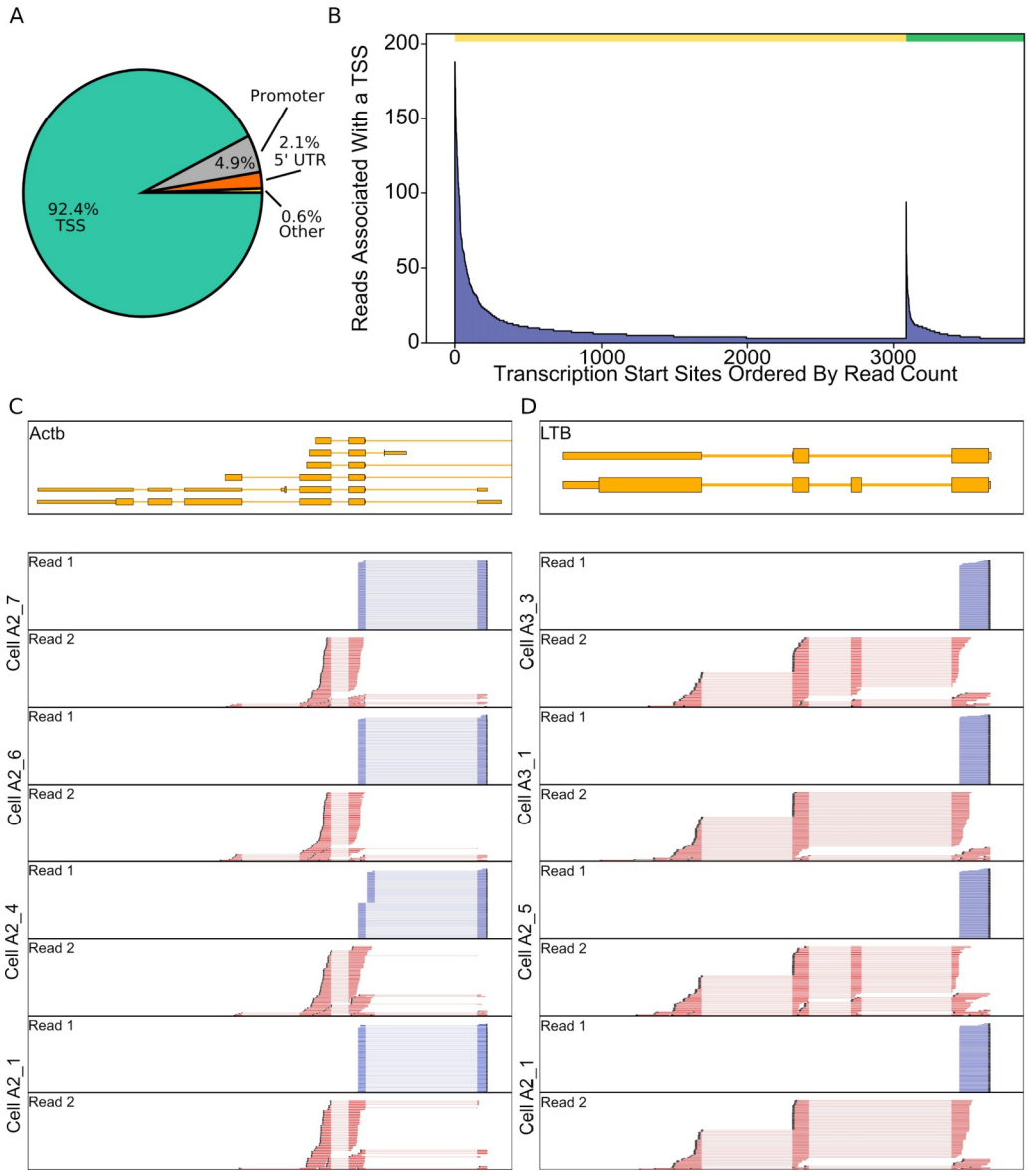

Fig 4. Transcription start sites are detected in single CD27^{high} CD38^{high} B cells
A) CD27^{high} CD38^{high} Tn5Prime peaks were matched to features in the Gencode annotation and the feature they matched are shown as a pie chart. TSS = on or less than 25bp behind the start of an annotated GENCODE gene, 5'UTR = inside 5' prime untranslated region, Promoter = between 26 and 1000bp behind start of annotated gene. B) Tn5 peaks were categorized into two groups. One group contains all peaks within 25bp of a peak identified in the complete RIKEN CAGE peak Human peak database and the other group contains all other peaks. These peaks were sorted by the number of cells associated with that peak in the CD27^{high} CD28^{high} B cell data set and displayed in figure 5a. The yellow bar indicates the peaks within 25bp and the green bar indicates all other peaks. C,D) Genome Browser view of

reads of several cells aligned to 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 not perturbed, 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 (Fig. S3). 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 dimensionality reduction. The clusters were robust when the data was subsampled to 100,000 reads per cell (Fig. S4). We then identified genes that showed significant differences between the two clusters. We detected 411 genes with significant changes including J-chain, LTB, XBP-1, HSPA5, MZB1, as well as HLA-DRA, HLA-DRB5, and HLA-DPB1 (Table S2). J-chain was upregulated in cluster 2 and is involved in antibody secretion of IgM and IgA (Lamson and Koshland 1984) (Fig. 5). We also found XBP-1, MZB1 and HSPA5 were upregulated within cluster 2 and are known targets of BLIMP-1 which is essential in plasmablast differentiation (Fig. S5) (Minnich et al. 2016). LTB was downregulated in cluster 2 and has been shown to be downregulated upon B cell activation (Zhu et al. 2004) (Fig. 5). HLA-DRA, HLA-DRB5, and HLA-DPB1 which encode for the alpha and beta chains of the MHC II complex were also

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 antibody and display less MHC II complex than the memory B cells in cluster 1.

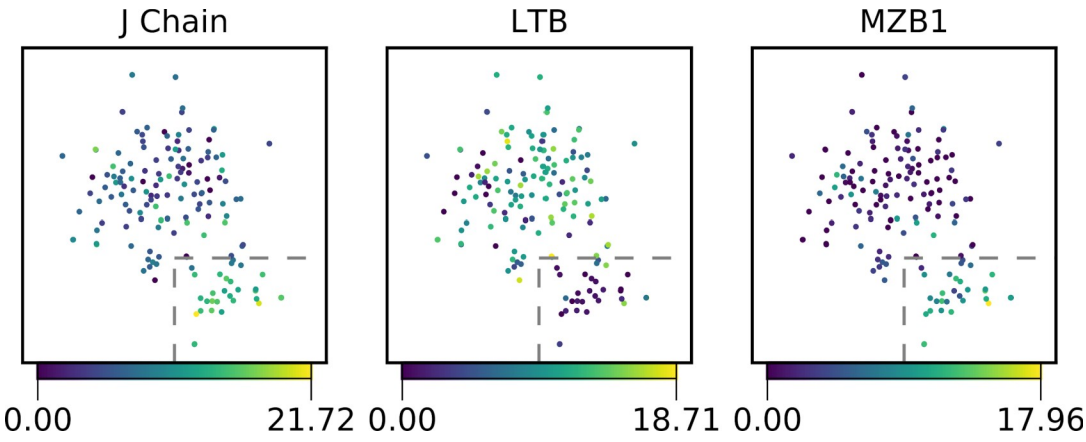

Figure 5. Clustering of CD27^{high} CD38^{high} B cells
159 B cells were divided into two populations by t-SNE dimensionality reduction (Maaten et al. 2008). In the three subplots, cells are colored based on their expression of example genes that were significantly differentially expressed between the 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 the sequences of the antibodies they express.

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). In contrast to 3' end based Drop-Seq and 10X Genomics data, 5' based Tn5Prime could potentially provide this 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 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. In this way, we effectively determined heavy and light chain sequences and identify their unique pairings within single B cells (Fig. 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 (Fig. S6). 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.

101 and 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.

First, the putative plasmablast population featured less IgM antibodies than the memory B cell population (19% IgM in plasmablasts vs 53% in memory B cells). Second, using IgBlast (Ye et al. 2013), we found that both heavy (Fig. 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). Third, 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. 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. 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.

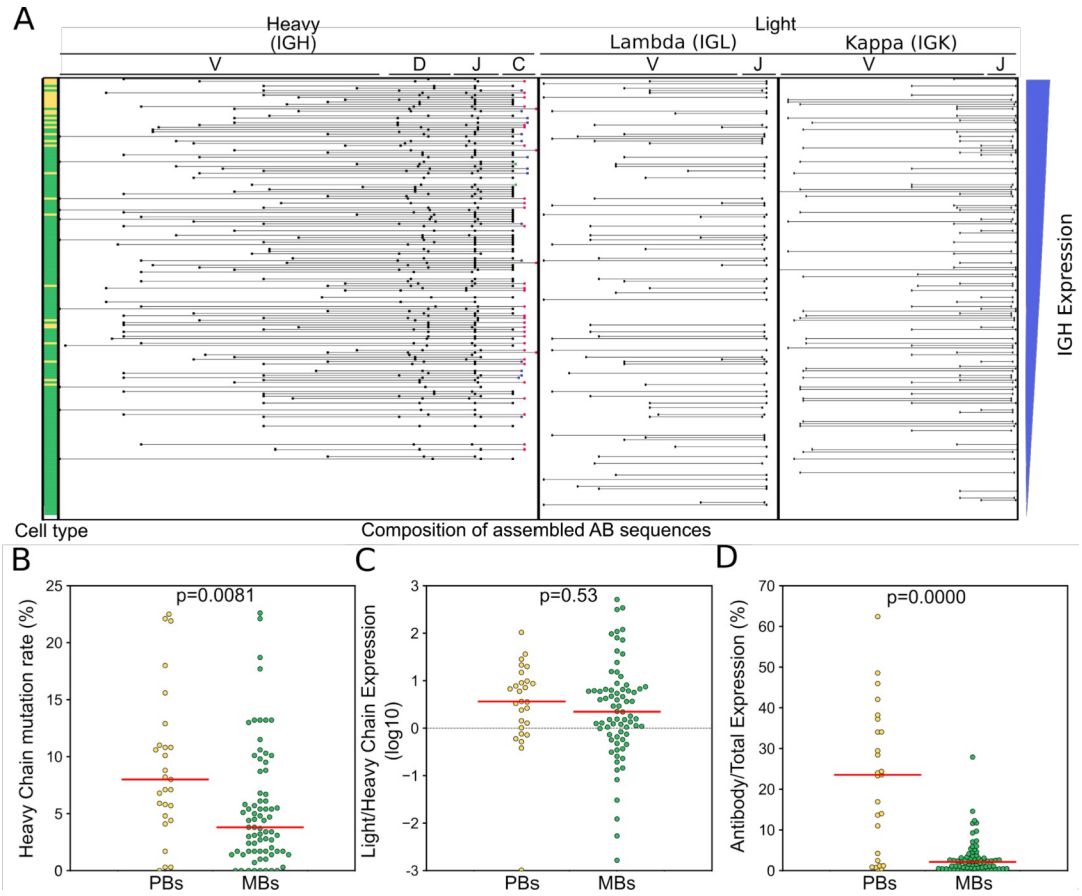

Figure 6. Assembling Antibody transcripts from Tn5Prime data

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 dotplots. 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 Smartseq2 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 order of magnitude more RNA. As the Smartseq2 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.

In contrast to other droplet or microwell based protocols, which interrogate only the 3' ends of transcripts, the Tn5Prime protocols 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.

To highlight the power of our approach we isolated 192 single human B cells from Peripheral Blood Mononuclear Cells(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 (X-box binding protein 1), J-chain, HSPA5, and MZB1

(Marginal Zone B1), 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 presents the currently highest

throughput mechanism to comprehensively analyze the individual cells of the adaptive immune system by determining both paired adaptive immune receptor sequences and gene expression profiles.

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 University of California, Santa Cruz (UCSC) vivarium according to Institutional Animal Care and Use Committee (IACUC)-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. Single human B cells were isolated from PBMC 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).

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

4ul of RNA or Single Cell Lysate was reverse transcribed using Smartscribe Reverse Transcriptase (Clontech) in a 10ul reaction including either a Smartseq2 TSO (Smartseq2 libraries) or a Nextera A TSO (Tn5Prime libraries) according to manufacturer's instructions at 42°C. 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 amplified with KAPA Hifi Readymix 2x (KAPA) using the ISPCR primer and a Nextera A Index primer (Tn5Prime only). The resulting PCR product was treated with Tn5 enzyme (Picelli, Björklund, et al. 2014) loaded with either Tn5ME-A/R and Tn5ME-B/R (Smartseq2) or Tn5ME-B/R adapters only (Tn5Prime).

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 Smartseq2 and Tn5Prime libraries were sequenced on an Illumina HiSeq2500 2x150 run, mouse B2 cell Tn5Prime libraries were sequenced on a Illumina MiSeq 2x300 run, and human B cell Tn5Prime libraries were sequenced on two Illumina HiSeq3000 runs.

Sequencing alignment and analysis

Smartseq2, 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) GM12878 data as well as Tn5Prime B2 data were trimmed of adapters low quality bases using trimmomatic (v0.33) (Bolger, Lohse, and Usadel 2014) and a quality cutoff of Q15. Trimming of the 192 human B cell data was performed by Cutadapt[Cutadapt](32), filtering out all paired reads where one or more read had a post-trimming length of less than 25 bp.

Trimmed reads derived from the GM cell line and 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 and Tn5Prime 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 TN5 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 TN5 peak coordinates. Each chromosome and orientation had its own kd-tree.

Antibody Assembly

After assignment, reads were assembled into transcriptomes using rnaSPAdes (Bankevich et al. 2012) with 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[BWA-MEM paper] (Li 2013).

Antibody transcripts were filtered with 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, the overall abundance(as determined by BWA) within the cell,the CDR3 sequence and the type(IGH,IGK,IGL) as well as the inferred segments used during VDJ recombination.
2. The transcripts are 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 counts added together. This is done to correct for the production of spurious alternative assemblies produced by SPADES within each cell's assembled transcriptome.

4. a list is generated for each transcript of the cells in which they appear.

5. The lists is sorted by the abundance of the transcript within the cells.

6. the entries in the lists are marked by their relative abundance. If the number of reads aligned to the transcript in a cell is less than 10% of the largest amount reads aligned to that transcript within any cell, it is marked as being a potential contaminant. The idea is that if a transcript discovered in a cell is a contaminant it should have at least an order of magnitude fewer reads associated with it when compared with the cell it actually came from.

7. For each type (IGH,IGK,IGL) of IG transcript found within each cell, the largest unique (non-contaminant) transcript is picked to have potentially come from that cell. if a unique transcript cannot be found, the most highly expressed transcript is selected

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

9. After this process, most cells should have a single heavy chain and a single 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) under the accessions

PRJNA320873 (GM12878 Smartseq2 and Tn5Prime), PRJNA320902 (Mouse B2 Cells), and

PRJNA415475 (Human $CD27^{high}$ $CD38^{high}$). 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 and CAGE Peak Caller and peak distance calculator are available at https://github.com/chkcole/Peak-Calling. All other Scripts are available upon request.

Acknowledgements

We thank the Sanford lab at UCSC for providing GM12878 cells and the Dubois lab at UCSC for expert help with producing Tn5 enzyme. This work was supported by an

NIH/NIDDK award (R01DK100917) and an Alex's Lemonade Stand Foundation Innovation Award to ECF; by CIRM Training grant TG2-01157 to AEB; and by CIRM Shared Stem Cell Facilities (CL1-00506) and CIRM Major Facilities (FA1-00617-1) awards to UCSC. ECF is the recipient of a California Institute for Regenerative Medicine (CIRM) New Faculty Award (RN1-00540) and an American Cancer Society Research Scholar Award (RSG-13-193-01-DDC). CV is a recipient of the 2017 Hellman Fellowship. A.B. and C.C. are funded by the NHGRI/NIH training grant 1T32HG008345-01.

Aim 3: Repertoire Sequencing Using the Oxford Nanopore

Abstract

Repertoire sequencing is a form of RNAseq specific to immunoglobulin transcripts. Over time, improvements have been made to this assay which have dramatically increased the breadth, depth, and overall accuracy of the resulting data. However, repertoire sequencing still requires a separate library preparation and sequencing steps independent of and in addition to the standard PolyA-specific RNAseq one might want to do on an RNA sample. We propose that repertoires can be acquired as a byproduct of a general RNA sequencing protocol by using the Rolling Circle to Concatemeric Consensus(R2C2) (Volden et al. 2018) method on lymphocyte-derived RNA. We show that the repertories derived from this method are comparable to specialized repertoire sequencing approaches and can be achieved at no additional cost to the experimenter.

Introduction

Antibodies are proteins generated exclusively by B cells whose primary purpose is to bind to pathogens, disabling them and marking them for destruction by the immune system. During B cell maturation, each antibody gene(Heavy, Kappa or Lambda) is produced through a process called V(D)J recombination whereby a single Variable, Diversity, and Joining gene segment is randomly selected and recombined, along with the addition of non-template nucleotides, to create a gene unique to that cell. This process is what allows the adaptive immune system to generate protection against nearly any pathogen it might encounter.

Before the advent of high-throughput sequencing, antibodies were characterized by Sanger sequencing of cloned antibody genes or cDNA. Cloning requires the isolation of a pure DNA template, so early studies were done almost exclusively on monoclonal B cell cancers. Even when all the necessary materials were present this method was slow and laborious. Indeed, some of the earliest studies of V segment diversity were the product of only a few dozen heavy or light chain sequences. With the invention of PCR and high-throughput sequencing came the ability to assay millions of discrete DNA molecules. In 2009 a method for sequencing antibody repertoires from zebrafish using the Roche 454 platform was published. Heavy chain Amplicons were generated from rna using primers specific to the first framing region of the V segment and the constant region. These amplicons were used as the template for a 2x230 bp run which, in zebrafish, were long enough to cover the entire V segment, allowing the entire heavy chain to be reconstituted through overlap

merging of the paired reads. Since then improvements to the protocol have been made, most notably the inclusion of UMIs into the read which result in the generation of higher-accuracy antibody transcript sequences as well as more accurate quantification of antibody expression. Nevertheless, amplicon-based repertoire sequencing has several experimental and informatic shortcomings such as the lack of an effective heavy/light chain pairing method in bulk samples, inability to distinguish between membrane and secreted antibodies, and the fact that the use of V and C segment primers can result in under or over-reporting of some antibodies. Two key technologies have emerged over the last five years which we believe will supplant dedicated repertoire-sequencing protocols, allowing for the production of high-quality repertoires as a byproduct of more general RNAseq experiments.

Nanopore Sequencing

Nanopore sequencing is a recently developed technology which permits the real-time sequencing of single molecules of DNA of an arbitrary length. This technology was pioneered by Oxford Nanopore Technologies (ONT) and is mostly widely implemented in the form of a device called the ONT MinION. Unlike previous sequencing technologies amplification of the template strand is not required in order to determine the sequence. Instead the DNA is threaded through a protein channel embedded in lipid membrane with a current running across it. The DNA strand interrupts the flow of electrons through the pore, and will do so in a manner which can be used to identify the exact sequence which happens to be inside the pore at any given time. Thus, by threading a strand of DNA and measuring the change in

current, the current trance can be used to infer its sequence. One of the drawbacks of this technology is that there is significantly more uncertainty about the correctness of each base in the sequence compared with older sequencing technologies. For example, Illumina reads have a typical error rate of 1 in 1000 bases while the Oxford Nanopore produces errors at a rate around 1 in 10. Depending on the downstream applications for this sequence, a high frequency of errors could make the data difficult to analyze or entirely unusable.

In 2018 our lab published a method for improving the accuracy for nanopore reads called Rolling Circle Amplification to Concatemeric Consensus (Volden et al. 2018). DNA fragments are circularized by gibson assembly and amplified using phi29 and random primers. This produces pieces of DNA whose sequences are concatemers of the original DNA fragment that was circularized. These pieces of DNA are sequenced and the sequence of the original fragments are inferred by identifying repeats in the sequence, splitting them, and then generating a consensus with the repeated sequences. Doing this we can generate sequences from the MinION with an average base error rate of 1 in 50.

The original motivation for the development of this method was transcript isoform analysis. Isoforms are transcripts produced by the same gene but with alternative splicing or transcription start sites. The choice of exons to include in a transcript can dramatically impact the function of the resulting protein. As such, the ability to quantify the relative expression of isoforms from the same gene is critical for

understanding a cell's biology. Short-read RNAseq can't be used to quantify isoforms from complex genes because they are not long enough to span all of the exons and provide the needed connectivity information. Oxford Nanopore solves that problem by allowing the sequencing of entire transcripts. However, the base accuracy can make aligning regular nanopore reads imprecise and may cause the analysis software to miss or misidentify isoforms. In addition, a 1 in 10 error rate makes multiplexing libraries impractical since most demultiplexing strategies assume Illumina-level error rates with no insertions or deletions in the barcode sequence. R2C2 provided a method for sequencing full-length transcripts at an error rate low enough to accurately call isoforms and demultiplex pooled single cell libraries.

We concluded that, with enough sequencing depth, R2C2 could be used to characterize antibody repertoires of B cell-containing samples without any steps in the library creation process. Indeed, we proposed these repertoires would be a byproduct of a non-selective RNA sequencing experiment. In order to test this hypothesis we generated libraries from Human PBMC-derived RNA (Sample 23_2) using R2C2 and a conventional antibody-repertoire sequencing protocol and compared the results. We also included previously published data using the TMI-seq approach as part of the analysis (Cole et al. 2016).

Results

R2C2

RNA was extracted from the PBMCs resulting in 235.6ng RNA /ul in 30ul of H2O. cDNA was generated from 4 ul of RNA using a Oligo-dT primer and template-switching oligo followed by ISPCR amplification for 12 cycles resulting in 8ng of cDNA/ul in 20ul. 40ng of cDNA was circularized by gibson assembly using DNA splints complementary to the ends of the cDNA molecule and the reactions were treated with a cocktail of exonuclease to remove uncircularized cDNA. The circles were SPRI cleaned and eluted in 40ul. The circularized cDNA was used as the input for four 50ul Rolling-Circle Amplification reactions using phi29 and random hexamers. These reactions were digested with T7 endonuclease and size selected on a gel to remove DNA less than 10kb long. This produced 65ng/ul in 100ul of H2O. This DNA was used as the input for four LSK109 ligation library preps (1ug per library) and sequenced on four flowcells using the Oxford Nanopore MinION, generating 16.6 million raw reads. Consensi were called from these reads as previously described (Volden et al. 2018), generating 10.3 million consensus reads. These were then aligned to the hg38 genome with all alternative assemblies removed using minimap2 using the splice-aware mode, aligning 99.96% of the reads (Li 2018). Putative antibody transcripts were collected by extracting reads which aligned on chromosome 14 between positions 105,536,018 and 106,883,485. This results in a set of 15171 reads of which 12864 could be successfully annotated by IGblast (Ye et al. 2013). V segment alleles and recombination rates were calculated for transcripts with either an IgM or IgD isotype which indicate that an antibody belongs to a naive B cell and is unlikely to contain significant amounts of somatic hypermutation.

Antibody Repertoires

The same pool of RNA used for the R2C2 experiments was used to sequence standard Heavy chain repertoire libraries (Vollmers et al. 2013). This produced [INSERT RAW PAIRED READS HERE] which were used to generate 49563 merged reads which covered the entire variable region and enough of the constant region to call isoforms. Of these, 48899 were successfully annotated by IgBlast.

V segment analysis

We determined the validity of our heavy chain sequences by examining the distribution of CDR3 lengths as well as the pattern of somatic hypermutations present in the V segments. CDR3 lengths follow a well established pattern: Their nucleotide length is generally a multiple of three and they are between 21 and 81 bases long. The CDR3 sequence is determined by locating the amino acid motifs, [FY] [FHVWY] C [ADEGIKMNRSTV] at the 3' end of the V segment and W[GAV] in the J segment. The location of these motifs determines the start and end of the CDR3. A histogram of the CDR3 lengths for the repertoire, TMIseq and R2C2 antibody reads can be seen in figure 1. They have a median CDR3 length of 42, 45, and 45 nucleotides respectively. Which is in agreement with previously published reports on CDR3 length for heavy chains.

In order to estimate the V segment recombination rate from the disparate data sets we only considered heavy chains with an IgM or IgD Isotype. The results of the

comparison can be seen in figure 2. The results indicate significant disagreement between the three different methods for the recombination rate of the V segments as well as the number and type of alleles. Even accounting for the higher error rate of the R2C2 method, there clearly exists V segment alleles which are dominant in the R2C2 data set but almost entirely absent in the other two. Much of this is likely due to the fact that the repertoire sequencing protocol uses a priming site in FR2 and will be unable to detect alleles with polymorphisms in FR1, CDR1, or the leader exon. In addition, it is likely that the primers will have different efficiencies for different V segments, possibly overrepresenting some recombination events while suppressing others. We see this same pattern in play when examining the isotype abundance in figure 3. In the R2C2, nearly 75% of all antibody transcripts are IGHA, while in the TMIseq data it's only a third. And, just like with the V segments, the TMI-seq and repertoire sequencing protocol prime off of the constant region in order to amplify antibody transcripts.

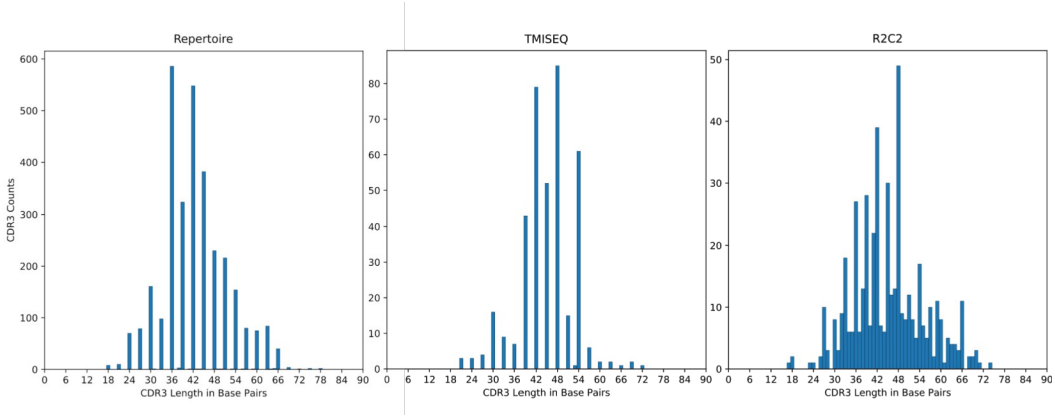

Fig 1. Histograms of the CDR3 lengths for antibody heavy chain transcripts detected in S23_2 RNA using three different methods: R2C2, TMIseq, and a conventional repertoire sequencing protocol.

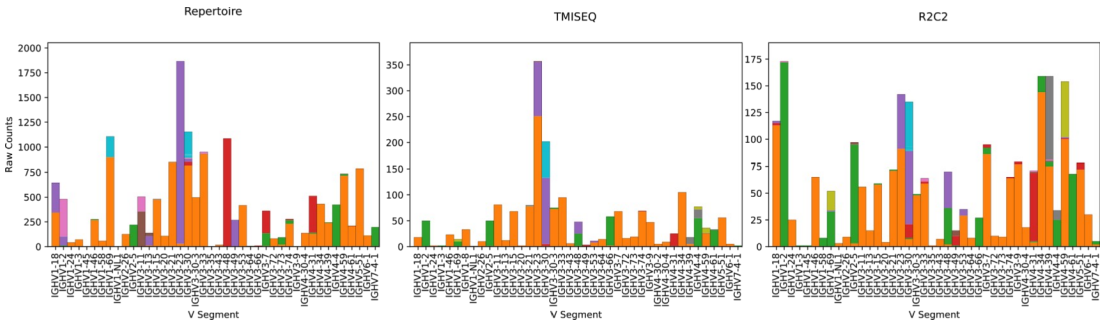

Figure 2. A bar chart comparing V segment recombination rates for R2C2, TMI-seq and a standard repertoire sequencing protocol. Bars are colored by allele.

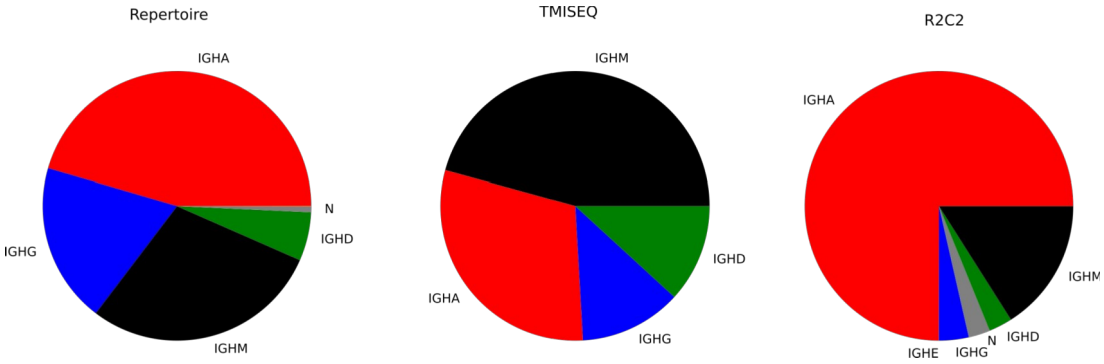

Figure 3. Pie chart of the heavy chain isotypes detected by R2C2, TMIseq, and a conventional repertoire sequencing protocol.

Methods

RNA Purification

Whole blood was collected from a fully consented human volunteer at the UCSC Student Health Center. PBMCs were isolated from Whole Blood Sample 23_2 by Ficoll-gradient and cryopreserved in DMSO(approximately 10 million live cells per tube) using liquid nitrogen (vapor phase). A single tube of PBMCs was defrosted and washed twice with 37C 1XPBS, spinning at 500g for 5 minutes in between washes.

500ul of trizol was added to the cell pellet and the whole volume was transferred to a 1.5ml tube. The cells were lysed for 5 minutes and 100ul of chloroform was added. The tube was shaken until the mixture acquired a pink, opaque hue. The reaction was allowed then allowed to sit for 2 minutes. The reaction was then spun at full speed in a microcentrifuge for 15 minutes at 4C. The aqueous layer was pipetted off and an equal volume of 70% Ethanol was added. The Qiagen RNAeasy mini prep kit was used, starting from section #2 in the instruction booklet, following the directions for purifying RNA. The RNA was eluted in 30ul of H2O, the concentration was measured with nanodrop, and aliquoted into individual PCR tubes and stored at -80C. The results from this procedure when tested on the nanodrop where 235.6ng of RNA per microliter of H2O.

cDNA

In order to generate enough cDNA for circularization we generated two separate reactions and then pooled them together at the end. First, 2ul of RNA at ~235.6ng/ul were combined with 1ul of unprotected oligo dT and 1 ul of dNTPs in a thin-walled PCR tube and heated to 72C for 5 minutes using a thermocycler and then immediately placed on ice. To this reaction, 2ul of 5x Smartscribe buffer, 1 ul DTT, 0.3ul of Smartscribe TSO, 1ul of Smartscribe enzyme, 0.25ul of Superase_IN, and 1.45ul of H2O were added. The reactions were then heated to 42C for 1 hour, and then 70C for 5 minutes. The two 10ul reactions were then combined and 1ul of 10um ISPCR(Smartseq2) primer, 1 ul of Lambda Exonuclease(NEB), 1ul RNAseA, and 25ul of 2X Kappa master mix were added for a total volume of 50ul. The reaction

was then underwent PCR(37C 30 minutes, 95C 0:30 seconds, 12 cycles: 98C 20S; 67C 15S; 72C 10 minutes, and then held for 4C). The reaction was then purified using a .85:1 SPRI purification ratio and eulted in 20ul of H2O.

Circularization

In order to circularize the cDNA we use double-stranded DNA splints complementary to the end of our cDNA that can undergo gibson assembly. In one PCR tube, 23ul of H2O, 25ul of Kappa 2X master mix, 1ul of UMI_Splint_ISPCR 100uM, and UMI_splint_reverse 100uM were added then cycled(95C 3 minutes, 98C 1 minute, 62C 1 minute, 72C 6 minutes). The reaction was then cleaned with the zymo select-a-size cleanup protocol for single-size selection, but in the buffer preparation add 85ul of 100% EtOH to 500ul of select-a-size DNA binding buffer. It was eluted in 20ul of buffer EB (elution buffer) and the concentration measured with qubit. For the gibson assembly reaction we combined an approximately equal mass of splint to cDNA. In a single PCR tube we added 0.2ul of Splint(210ng), 5ul of cDNA(8ng/ul), 4.8ul of H2O, and 10ul of NEBuilder 2X Master Max. This reaction is then incubated for 50C for 60 minutes. To this reaction we added 5ul of NEBuffer 2, 16ul H2O, 3ul Exonuclease I, 3ul of Exonuclease III, and 3ul of Lambda Exonuclease. This reaction is then incubated 37C for 16hr followed by a heat inactivation step of 80C for 20 minutes. The reaction was then purified with 0.8:1 SPRI beads and eluted in 40ul of H2O. in order to generate enough DNA for multiple library preps, we created four separate RCA reactions which were then combined at the end. For each reaction we combined in a PCR tube 5ul of Phi29 Buffer(NEB, 10X), 1ul of Phi29 Polymerase, 2.5ul dNTP(10nM), 2.5ul Random hexamer primers(exo resistant, 10uM), 10ul of

circularized DNA from the previous reaction, and 29ul of H2O. The tube was incubated for 30C overnight. The four 50ul reactions were combined into two 100ul reactions. To each tube 2.5ul of T7 was added and the tube incubated at 37C for 2 hours, agitating occasionally. The reactions were SPRI purified using a 0.5:1 ratio and eluted in 50ul of H2O. The DNA concentration was measured as 65ng/ul using qubit.

R2C2 Sequencing:

One microgram of size-selected and purified DNA was used as the basis for a 1D ligation-based library prep and then sequenced for 48 hours using the Oxford Nanopore MinION. This was done 4 times to produce the data set of interest.

Repertoires

In a single PCR tube, 1ul of RNA at 235ng/ul, 8.5ul of H2O, 2ul of IgH_RT primer mix, and 1ul of dNTP were added and heated to 72C for 5 minutes and then immediately placed on ice. To this reaction, 4ul of 5X Smartscribe buffer, 2ul of DTT, 1ul of Superase_IN, and 0.5ul of Smartscribe polymerase were added. The reaction was then heated to 42C for 1 hour followed by 70C for 5 minutes. To this reaction was added 10ul 5X Phusion buffer, 1ul IgH_C_pool nm, 1ul IgH_V_pool nm, 1ul dNTPs 10um, 15ul H20, 1.5ul DMSO, and 0.25ul Phusion Polymerase. The reaction was then cycled(2 cycles of: 98 °C for 3 min, 52 °C for 2 min, 72 °C for 4 min) and then purified using a Zymo Select-A-Size Clean and Concentrator column with a 300bp cutoff. The DNA was eluted in 23ul of buffer EB.In a single PCR tube was added the 23ul of cDNA, 25ul of 2X Kappa Master Mix, 1ul of Nextera A primers

(10nM), and 1ul of Nextera B primers (10nM). The reaction was cycled (95°C for 3 minutes, 25 cycles of: 98°C 20 secs; 63°C for 30 secs;72°C for 2 min, final extension of 72°C for 5 minutes). A final purification step was performed using a 1:0.75 sample to SPRI bead purification and the product was eluted in a final volume of 20uL. The sample was then ran on a 2% agarose gel and quantified by Qubit fluorometer (Invitrogen) to verify quantity and quality of the library prior to sequencing. The samples were sequenced on a MiSeq 2 x 300 run.

Conclusion

During the course of my studies I tackled several problems in the field of antibody genetics and succeeded in solving two of them. The TMIseq method allows the assembly of full-length antibody transcripts from molecularly-index subreads. This method solved the issue of full-length transcript sequencing, allowing people to accurately assign V segment allele and isotype to individual transcripts. The Tn5prime method is a simple method for sequencing the 5' ends of RNA transcripts. In addition, it allows people to pair full-length heavy and light chains in individual B cells. I demonstrated that the R2C2 method can be used to sequence heavy and light chain transcripts with enough accuracy to confidently assign V segment alleles and isotypes. The preliminary data indicates substantial disagreement with previously published methods, raising questions about the validity of V segment

recombination rate estimation from primer-based assays.

I propose that the R2C2 method could one day be used as a substitute for repertoire-specific sequencing protocols when assaying immune function. It captures full-length transcript information for antibodies, T cell receptors, and every other polyA transcript present in the sample. It is also accurate enough to assign transcripts to specific isoforms and alleles and, as such, is a substantially more powerful method than assays which use primers to target only specific transcripts. I also propose that because the R2C2 method does not introduce bias into the data by using DNA primers that we can accurately measure the true recombination rates for specific gene segments, perhaps revealing hitherto unknown biology about VDJ recombination.

References

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.

Arnaout, Ramy, William Lee, Patrick Cahill, Tracey Honan, Todd Sparrow, Michael Weiand, Chad Nusbaum, Klaus Rajewsky, and Sergei B. Koralov. 2011. “High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans.” *PloS One* 6 (8): e22365.

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.

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

Boyd, Scott D., Eleanor L. Marshall, Jason D. Merker, Jay M. Maniar, Lyndon N. Zhang, Bita Sahaf, Carol D. Jones, et al. 2009. “Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel VDJ Pyrosequencing.” *Science Translational Medicine* 1 (12): 12ra23.

Brack, C., M. Hirama, R. Lenhard-Schuller, and S. Tonegawa. 1978. “A Complete Immunoglobulin Gene Is Created by Somatic Recombination.” *Cell* 15 (1): 1–14.

Brack, C., and S. Tonegawa. 1977. “Variable and Constant Parts of the Immunoglobulin Light Chain Gene of a Mouse Myeloma Cell Are 1250 Nontranslated Bases Apart.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5652–56.

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.

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* 46 (10): e62.

Cole, Charles, Roger Volden, Sumedha Dharmadhikari, Camille Scelfo-Dalbey, and Christopher Vollmers. 2016. “Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier–Guided Amplicon Assembly.” *The Journal of Immunology* 196 (6): 2902–7.

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.

DeKosky, Brandon J., Takaaki Kojima, Alexa Rodin, Wissam Charab, Gregory C. Ippolito, Andrew D. Ellington, and George Georgiou. 2015. “In-Depth Determination and Analysis of the Human Paired Heavy- and Light-Chain Antibody Repertoire.” *Nature*

Medicine 21 (1): 86–91.

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.

Early, P., H. Huang, M. Davis, K. Calame, and L. Hood. 1980. "An Immunoglobulin Heavy Chain Variable Region Gene Is Generated from Three Segments of DNA: VH, D and JH." *Cell* 19 (4): 981–92.

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

Georgiou, George, Gregory C. Ippolito, John Beausang, Christian E. Busse, Hedda Wardemann, and Stephen R. Quake. 2014. "The Promise and Challenge of High-Throughput Sequencing of the Antibody Repertoire." *Nature Biotechnology* 32 (2): 158–68.

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.

Glanville, J., T. C. Kuo, H. -. C. von Budingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, et al. 2011. "Naive Antibody Gene-Segment Frequencies Are Heritable and Unaltered by Chronic Lymphocyte Ablation." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1107498108.

He, Linling, Devin Sok, Parisa Azadnia, Jessica Hsueh, Elise Landais, Melissa Simek, Wayne C. Koff, Pascal Poignard, Dennis R. Burton, and Jiang Zhu. 2014. "Toward a More

Accurate View of Human B-Cell Repertoire by next-Generation Sequencing, Unbiased Repertoire Capture and Single-Molecule Barcoding." *Scientific Reports* 4 (October): 6778.

Hiatt, Joseph B., Rupali P. Patwardhan, Emily H. Turner, Choli Lee, and Jay Shendure. 2010. "Parallel, Tag-Directed Assembly of Locally Derived Short Sequence Reads." *Nature Methods* 7 (2): 119–22.

Hong, Lewis Z., Shuzhen Hong, Han Teng Wong, Pauline P. K. Aw, Yan Cheng, Andreas Wilm, Paola F. de Sessions, et al. 2014. "BAsE-Seq: A Method for Obtaining Long Viral Haplotypes from Short Sequence Reads." *Genome Biology* 15 (11): 517.

Honjo, T., and T. Kataoka. 1978. "Organization of Immunoglobulin Heavy Chain Genes and Allelic Deletion Model." *Proceedings of the National Academy of Sciences of the United States of America* 75 (5): 2140–44.

Howie, Bryan, Anna M. Sherwood, Ashley D. Berkebile, Jan Berka, Ryan O. Emerson, David W. Williamson, Ilan Kirsch, et al. 2015. "High-Throughput Pairing of T Cell Receptor α and β Sequences." *Science Translational Medicine* 7 (301): 301ra131.

Hozumi, N., and S. Tonegawa. 1976. "Evidence for Somatic Rearrangement of Immunoglobulin Genes Coding for Variable and Constant Regions." *Proceedings of the National Academy of Sciences of the United States of America* 73 (10): 3628–32.

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

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.

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.

Jiang, Ning, Jiankui He, Joshua A. Weinstein, Lolita Penland, Sanae Sasaki, Xiao-Song He, Cornelia L. Dekker, et al. 2013. “Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination.” *Science Translational Medicine* 5 (171): 171ra19.

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

Kidd, Marie J., Zhiliang Chen, Yan Wang, Katherine J. Jackson, Lyndon Zhang, Scott D. Boyd, Andrew Z. Fire, Mark M. Tanaka, Bruno A. Gaëta, and Andrew M. Collins. 2012. “The Inference of Phased Haplotypes for the Immunoglobulin H Chain V Region Gene Loci by Analysis of VDJ Gene Rearrangements.” *Journal of Immunology* 188 (3): 1333–40.

Komori, T., A. Okada, V. Stewart, and F. W. Alt. 1993. “Lack of N Regions in Antigen Receptor Variable Region Genes of TdT-Deficient Lymphocytes.” *Science* 261 (5125): 1171–75.

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.

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

———. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100.

Logan, A. C., H. Gao, C. Wang, B. Sahaf, C. D. Jones, E. L. Marshall, I. Buno, et al. 2011. "High-Throughput VDJ Sequencing for Quantification of Minimal Residual Disease in Chronic Lymphocytic Leukemia and Immune Reconstitution Assessment." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1118357109.

Looney, Timothy J., Ji-Yeun Lee, Krishna M. Roskin, Ramona A. Hoh, Jasmine King, Jacob Glanville, Yi Liu, et al. 2016. "Human B-Cell Isotype Switching Origins of IgE." *The Journal of Allergy and Clinical Immunology* 137 (2): 579–86.e7.

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.

Lundin, Sverker, Joel Gruselius, Björn Nystedt, Preben Lexow, Max Käller, and Joakim Lundeberg. 2013. “Hierarchical Molecular Tagging to Resolve Long Continuous Sequences by Massively Parallel Sequencing.” *Scientific Reports* 3: 1186.

Matsuda, F., K. Ishii, P. Bourvagnet, K. i. Kuma, H. Hayashida, T. Miyata, and T. Honjo. 1998. “The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus.” *The Journal of Experimental Medicine* 188 (11): 2151–62.

Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research*9.Nov (2008): 2579-2605.

McKean, D. J., M. Bell, and M. Potter. 1978. “Mechanisms of Antibody Diversity: Multiple Genes Encode Structurally Related Mouse Kappa Variable Regions.” *Proceedings of the National Academy of Sciences of the United States of America* 75 (8): 3913–17.

Meyer, Everett H., Andro R. Hsu, Joanna Liliental, Andrea Löhr, Mareike Florek, James L. Zehnder, Sam Strober, et al. 2013. “A Distinct Evolution of the T-Cell Repertoire Categorizes Treatment Refractory Gastrointestinal Acute Graft-versus-Host Disease.” *Blood* 121 (24): 4955–62.

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.

Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo. 2000. “Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme.” *Cell* 102 (5): 553–63.

Oettinger, M. A., D. G. Schatz, C. Gorka, and D. Baltimore. 1990. “RAG-1 and RAG-2,

Adjacent Genes That Synergistically Activate V(D)J Recombination." *Science* 248 (4962): 1517–23.

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

Pernis, B., L. Forni, and A. L. Luzzati. 1977. "Synthesis of Multiple Immunoglobulin Classes by Single Lymphocytes." *Cold Spring Harbor Symposia on Quantitative Biology* 41 Pt 1: 175–83.

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.

Rossano, Joseph W., Susan W. Denfield, Jeffrey J. Kim, Jack F. Price, John L. Jefferies, Jamie A. Decker, E. O'brian Smith, Sarah K. Clunie, Jeffrey A. Towbin, and William J. Dreyer. 2009. "Assessment of the Cylex ImmuKnow Cell Function Assay in Pediatric Heart Transplant Patients." *The Journal of Heart and Lung Transplantation: The Official Publication of the International Society for Heart Transplantation* 28 (1): 26–31.

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.

Schatz, D. G., M. A. Oettinger, and D. Baltimore. 1989. "The V(D)J Recombination

Activating Gene, RAG-1." *Cell* 59 (6): 1035–48.

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.

Shugay, Mikhail, Olga V. Britanova, Ekaterina M. Merzlyak, Maria A. Turchaninova, Ilgar Z. Mamedov, Timur R. Tuganbaev, Dmitriy A. Bolotin, et al. 2014. "Towards Error-Free Profiling of Immune Repertoires." *Nature Methods* 11 (6): 653–55.

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.

Tonegawa, S. 1983. "Somatic Generation of Antibody Diversity." *Nature* 302 (5909): 575–81.

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.

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.

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, Iwijn De Vlaminck, Hannah A. Valantine, Lolita Penland, Helen Luikart, Calvin Strehl, Garrett Cohen, Kiran K. Khush, and Stephen R. Quake. 2015. “Monitoring Pharmacologically Induced Immunosuppression by Immune Repertoire Sequencing to Detect Acute Allograft Rejection in Heart Transplant Patients: A Proof-of-Concept Diagnostic Accuracy Study.” *PLoS Medicine* 12 (10): e1001890.

Vollmers, Christopher, Rene V. Sit, Joshua A. Weinstein, Cornelia L. Dekker, and Stephen R. Quake. 2013. “Genetic Measurement of Memory B-Cell Recall Using Antibody Repertoire Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (33): 13463–68.

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.

Weigert, M. G., I. M. Cesari, S. J. Yonkovich, and M. Cohn. 1970. “Variability in the Lambda Light Chain Sequences of Mouse Antibody.” *Nature* 228 (5276): 1045–47.

Weinstein, Joshua A., Ning Jiang, Richard A. White 3rd, Daniel S. Fisher, and Stephen R. Quake. 2009. “High-Throughput Sequencing of the Zebrafish Antibody Repertoire.” *Science* 324 (5928): 807–10.

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, Nicholas C., Justin De La Cruz, Laith Q. Al-Mawsawi, C. Anders Olson, Hangfei Qi, Harding H. Luan, Nguyen Nguyen, et al. 2014. "HIV-1 Quasispecies Delineation by Tag Linkage Deep Sequencing." *PloS One* 9 (5): e97505.

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.

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.

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.