THE UNIVERSITY OF CHICAGO

THE DEVELOPMENT OF SPEC-SEQ AND ITS APPLICATION TO PERIPHERAL BLOOD PLASMABLASTS AFTER INFLUENZA CHALLENGE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON IMMUNOLOGY

BY
KARLYNN NEU DULBERGER

CHICAGO, ILLINOIS
DECEMBER 2017

Table of Contents

List of Figures

List of Supplementary Figures (Supplementary files available online)

List of Tables

List of Supplementary Tables (Supplementary files available online)

Acknowledgements

Overall, I found my graduate experience very challenging but also transformative and I am very happy and proud to be completing it. I would like to thank everyone who was a part of my life over the last 6 years, each conversation was taken to heart and in your own way you all helped me get here. Additionally, I would like to take this occasion to thank a few people in particular.

First of all, I would like to thank Patrick C. Wilson, my thesis mentor. Not only would this entire project have been impossible without his boundless knowledge of everything B cells and his endless project ideas, but I also never would have come this far without his support, encouragement and enthusiasm. He genuinely believed that I could do anything and that all my hard work would pay off in the end. Aside from his excellence as a mentor he also has established a truly remarkable lab. From the very beginning of my thesis work I was surrounded by some of the most skilled and collaborative people I have ever worked with. I would like to thank all current and past members of the Wilson lab for their support. In particular, I would like to thank Carole Henry who is the incredible person and scientist who taught me everything I needed to know in the Wilson lab and who was always there to pick me up when things got tough. Also, Angela Hayes, Gloria Davis and Karla Rojas for always keeping things warm and interesting on the third floor of BSLC.

I would also like to take this opportunity to thank Aly A. Khan, who taught me everything I know about analyzing RNA-sequencing data. I doubt either of us knew how integral he would be for my thesis work when he agreed to be on my committee. Thank

you for being my off the books ‘second mentor’, for the many hours we spent solving problems and for the many opportunities and skills you have given me.

I am also thankful to the rest of my thesis committee for their time providing key guidance and support. Thank you Maria-Louisa Alegre for always being a sound voice in our meetings by focusing on the big picture aspects of my project and reminding my ambitious committee that I did have to graduate eventually. Thank you Fotini Gounari for teaching me about gene set enrichment analysis and introducing me to Metascape. Thank you Marcus Clark for always pushing me to do my best.

Outside of my scientific support system I would like to thank all of my family and friends for providing their love and understanding throughout this process. To my dad, Gary Neu, thank you for the many hours on the phone helping me keep science in perspective with life. To my mom, Rebekha Neu, who has had a life-long love for plasma cells, thank you for your unending support. Also, a special thank you to my Hyde Park family unit. Parker is always around for a cuddle and reminds me to be more like a cat. Nancy Levner is one of the most inspiring people I know and she always has a helpful bit of wisdom to share right before doing something decidedly unwise. Jen La’s fervor for life is contagious and her ever-supportive friendship will be one of the things I cherish most from the University of Chicago. Finally, a huge thank you to my husband, Charlie Dulberger, whose love for science is an inspiration to me. I don’t think either of us had any idea what we were getting into 7 years ago when we decided; ‘let’s just apply to graduate school’, but I am so glad to have come on this adventure with you.

Abstract

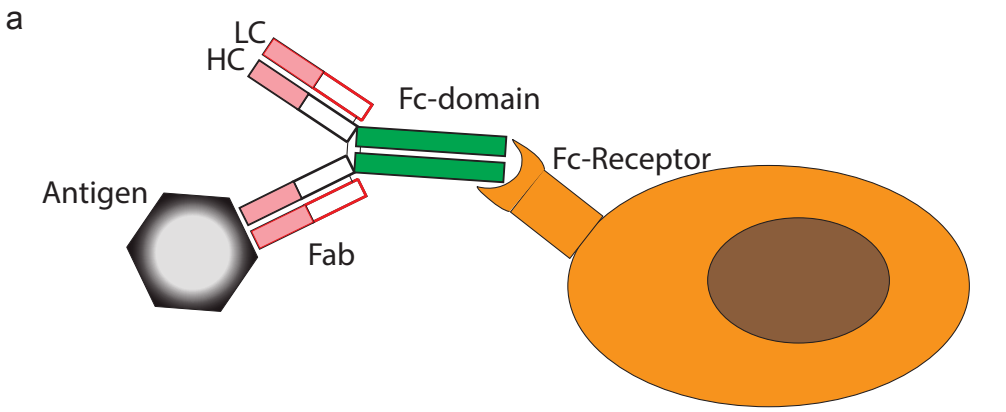
B cells are an integral part of the adaptive immune system and can differentiate into antibody secreting cells (ASC) in order to mediate a humoral immune response. Antibodies are soluble forms of the B cell receptor (BCR) that can bind antigen as well as elicit specific immune responses with their exchangeable constant domain. Single cell studies have extensively characterized the BCR sequence and functional capacity, but linking this to the original B cell gene expression profile has remained out of reach. To address this, we developed Spec-seq, an approach that facilitates simultaneous transcriptome and repertoire characterization at the single cell level. The unique dataset provided through this framework allowed for the development of BASIC, an algorithm to identify full-length BCR sequences from single cell RNA-sequencing data. We applied Spec-seq, and BASIC, to human peripheral blood plasmablasts seven days after influenza vaccination. We identified immunoglobulin genes as the major driver of plasmablast subpopulations, and also found differentially expressed genes between plasmablasts expressing vaccine-binding BCRs and those with vaccine-negative BCRs. The most significant of these genes have been credited in the literature with regulating antibody glycosylation, which can alter the immunogenicity of the secreted antibodies. Furthermore, we found that B cells of a single clonotype are more transcriptional similar than unrelated B cells. Overall, our findings have exposed previously unappreciated heterogeneity within plasmablast populations that are commonly utilized for the development of clinically relevant therapeutic antibodies and our Spec-seq framework will facilitate new avenues for the study of adaptive immune cells.

1. Introduction

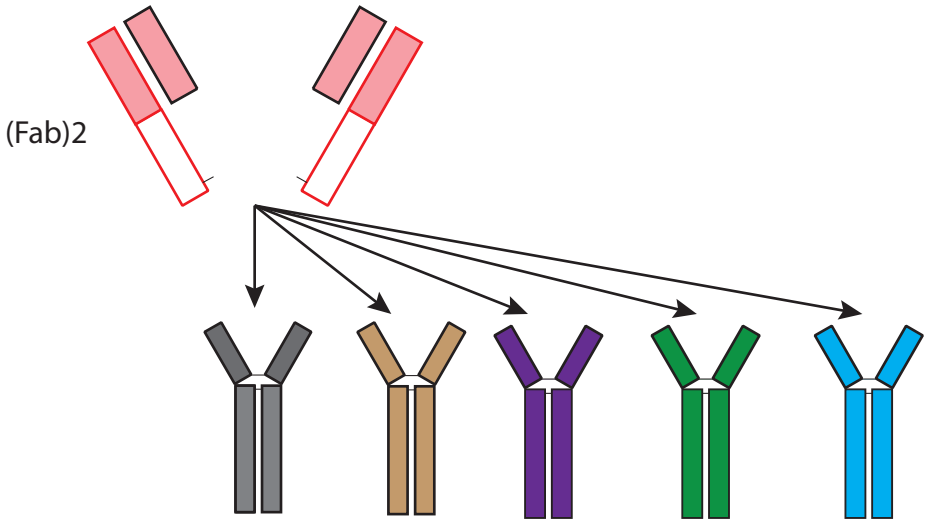
1.1 The humoral immune system

The immune system is designed to protect us from disease, whether it be from internal or external sources. There are two main branches of the immune system, the innate and the adaptive branches. Each branch is composed of immune cell types that are defined by their function as well as the type of antigen receptor they express on the cell surface. Innate immune cells have invariant receptors and comprise the first line of defense against challenge, whereas adaptive immune cells have a unique somatically-rearranged antigen receptor that is only shared with clonally-related progeny of the parent cell (Burnet, 1976). These receptors are generated through a complex process called V(D)J recombination, where a wide array of possible V, D and J genes are selected and recombined at the DNA level (Tonegawa, 1983). Both B and T cells utilize these adaptive receptors for antigen recognition, but in different ways. The T cell receptor (TCR) recognizes antigens, such as peptides, presented within major histocompatibility complex proteins (MHC), whereas the B cell receptor (BCR) recognizes antigens as intact protein epitopes (Murphy and Weaver, 2016) (Figure 1.1a). The BCR can also be made in a secreted form, known as an antibody.

Antibodies are the effector molecules of a humoral immune response, and are the basis for seven of the eleven nobel prizes awarded for discoveries in the field of immunology since 1901 (Nobelprize, 2010). The medical field has been harnessing their protective capabilities to improve human health since the original smallpox variolation (Jenner, 1798), long before antibodies were identified (Edelman, 1972; Ehrlich, 1902; Porter, 1972). Antibodies have since been branded as the first line of protection induced



b



Constant Domain	IgM	IgD	IgE	IgG (1-4)	IgA (1 & 2)
ADCC	-	-	-	++/-/++/-	-
CDC	+++	-	-	++/+/+++/-	+
ADCP	-	-	-	++/+/++/+	+

Figure 1.1 The B cell receptor. (**a**) The BCR is a Y shaped protein dimer with a heavy (outlined in black) and light chain (outlined in red). The top of each antibody contains two Fab (fragment-antigen binding) portions which contail the variable regions (colored light pink) and mediate antigen reactivity. The heavy chain constant domain makes up the base of the receptor and is also known as the Fc-domain because this region can be bound by Fc receptors on other immune cells. (**b**)The constant domain can be swapped through class switch recombination (CSR) to give the secreted BCR, or antibody, distinct effector functions. The relative capacity for three of the major functions are indicated; Antibody dependent cellular cytotoxicity (ADCC), complement dependent cellular cytotoxicity (CDC) and antibody dependent cellular phagocytosis (ADCP). Part b of this figure is a reproduction from Janeway's Immunobiology text book by Kenneth Murphy.

by vaccines, and the main mediators of ‘serum therapy’, where immunity can be attained in recipients of serum transferred from patients previously exposed to the antigen (von Behring and Kitasato, 1890). This has been a successful treatment against bacteria, toxins and infectious diseases for almost 130 years. Antibodies experienced a further increase in popularity when the ability to generate antibodies with a single specificity, known as monoclonal antibodies (mAb), was developed (Köhler and Milstein, 1975). Therapeutic mAb have had remarkable successes against infectious diseases (Beeler and van Wyke Coelingh, 1989; Young, 2002), autoimmunity (Chan and Carter, 2010), organ transplantation (Emmons and Hunsicker, 1987) and cancer (Grillo-López et al., 1999). Their popularity grows each year as more mAb therapeutics are licensed for use (Beck et al., 2017; FDA, 2017).

To induce an antibody response activated B cells must differentiate into antibody-secreting cells (ASCs). B cell activation occurs when a BCR binds to an antigen, which triggers the endocytosis of both the antigen and the BCR, and the complex is processed into peptides for antigen presentation in MHC class II molecules. A T cell that has acquired specificity for the same antigen can then recognize this MHC-peptide complex on the surface of the B cell and form an immunological synapse. This B-T cell interaction triggers critical signaling pathways that initiate entry of the B cell into a germinal center. A germinal center is a well-defined focus of cells located within lymphoid-associated tissues (originally identified by Walther Flemming in 1884) where B cells undergo affinity maturation, class switch recombination (CSR) (Lawton et al., 1975) and differentiation into ASCs (Butcher et al., 1982; reviewed in Nieuwenhuis and Opstelten, 1984).

Once a B cell enters a germinal center, affinity maturation begins, whereby a B cell undergoes selection for improved receptor binding affinity against a specific antigen (Eisen and Siskind, 1964). This is accomplished through alternating rounds of BCR somatic hypermutation (SHM) (McKean et al., 1984; Rajewsky et al., 1987; Weigert et al., 1970) and proliferation, with intermittent entry into the light zone of the germinal center where the ability of the mutated BCR to capture intact antigen off the surface of follicular dendritic cells (FDC) is tested (Lindhout et al., 1993). Successful capture is assessed by an additional B-T cell interaction, where T follicular helper cells give survival signals to B cells still successfully presenting antigen via MHC-II molecules (Victora et al., 2010). This interaction can either direct re-entry into the germinal center for additional rounds of affinity maturation, or graduation from the germinal center reaction and differentiation into different effector populations. At this step, the T cell can also provide signals via cytokines that direct CSR (Berton et al., 1989; Islam et al., 1991; Lundgren et al., 1989; Snapper and Paul, 1987; reviewed in Stavnezer, 1996) where the BCR heavy chain constant domain is exchanged with that of another isotype (Kataoka et al., 1980; Nossal et al., 1964; Rabbitts et al., 1980) (Figure 1.1b).

CSR is an important step before antibody secretion, as the BCR constant domain is a key modulator of effector capacity during a humoral immune response (Winkelhake, 1978). There are nine different constant domains to choose from. IgM is the first isotype expressed by activated B cells and is capable of neutralization, but primarily functions in activating complement in the blood or lymph (Brüggemann et al., 1987). The IgG group contains four subtypes with slightly varied functions and is the main isotype of antibody found in peripheral blood serum (Gonzalez-Quintela et al., 2008). IgG1 and IgG3

primarily function to activate an immune response through antibody-dependent cell-mediated cytotoxicity (ADCC) or complement-dependent cytotoxicity (CDC) (Brüggemann et al., 1987). ADCC is initiated when antibodies bound to target cells are recognized by Fc-receptors on cytolytic cells, whereas in CDC, the C1q protein binds to the Fc portion of the antibody and membrane attack complexes are generated (Murphy and Weaver, 2016; Schifferli et al., 1986). These features are desirable in an effective anti-viral response and the increased stability of IgG1 makes it an attractive target for vaccine and therapeutic antibodies (Brüggemann et al., 1987; Wrammert et al., 2008; Xu et al., 2012). IgA antibodies (IgA1 and IgA2) operate mainly through neutralization or immune exclusion and are the primary antibodies detected in mucosal secretions (Kerr, 1990). These characteristics of IgA antibodies make them highly effective at preventing pathogen entry into a host, which is a key objective for current anti-viral vaccine design (Chen and Cerutti, 2010; Neutra and Kozlowski, 2006; van Riet et al., 2012; Rose, 2014).

The differential glycosylation of antibody Fc-domains adds a further level of control over the cellular immune responses directed by antibodies. The CH2 domain of the constant region of all IgG and IgA BCR subtypes contains a conserved glycan, which can have different moieties added to direct Fc-effector functions (Hayes et al., 2014). The addition of a fucose moiety at the glycan base has been shown to reduce IgG affinity for the FcγRIIIa, which limits antibody immunogenicity by reducing ADCC (Shields et al., 2002; Yamane-Ohnuki et al., 2004). This modification is common on ‘steady-state’ antibodies of both IgG and IgA isotypes (Gomes et al., 2008; Shields et al., 2002), although it remains unclear how this modification specifically alters IgA Fc-

receptor binding (Shade and Anthony, 2013; Woof and Mestecky, 2005). The addition of terminal galactose moieties to the core glycan increases the ability of IgG antibodies to facilitate CDC (Hodoniczky et al., 2005), while this modification on IgA antibodies enhances ADCC through improved binding affinity with the FcαRI (Basset et al., 1999). Galactose modifications increase after influenza vaccination within the antigen-specific serum antibody compartment when compared to the steady-state serum profile (Selman et al., 2012). A concordant decrease in the frequency of fucose modified antibodies is also detected after vaccination (Selman et al., 2012). In combination, this data suggests fucosylated antibodies are common in steady-state serum where they display anti-inflammatory properties. In contrast, the galactose-modified antibodies are antigen induced, antigen specific and direct pro-inflammatory cellular responses (Shade and Anthony, 2013).

The addition of these modifications has been directly attributed to two major glycotransferase enzymes; *FUT8* and *B4GALT1*. The *FUT8* enzyme is required for the addition of fucose moieties to the BCR glycans (Yamane-Ohnuki et al., 2004), whereas the *B4GALT1* enzyme is necessary for the addition of galactose moieties (Keusch et al., 1998). The relative expression of these enzymes within bulk populations has been used as a proxy to assess differential glycosylation (Wang et al., 2015). Since controlling the antibody glycosylation profile is critical for monoclonal therapeutic development, cell lines lacking these enzymes are being developed (Yamane-Ohnuki et al., 2004). Furthering our understanding of the impact, and regulation, of antibody differential glycosylation is incredibly important for human health, especially as increasing numbers of monoclonal antibody therapeutics continue to be developed.

1.2 Influenza and the influenza vaccine

Seasonal influenza infection causes three to five million people to become severely ill each year, with around 10% of the cases resulting in death (WHO, 2016). Influenza is an RNA virus with many animal reservoirs (Medina and García-Sastre, 2011), making the eradication of this virus very challenging. Multiple genera of the influenza virus exist, with types A and B comprising the main health risks for humans. Within the influenza A and B virus types, there are numerous strains differentiated by their surface expression of two key proteins; hemagglutinin (HA) and neuraminidase (NA) (Laver and Valentine, 1969; Neu et al., 2016).

HA is responsible for viral adhesion to sialic acid receptors on host cells and subsequent membrane fusion (Bullough et al., 1994; Skehel et al., 1982; Weis et al., 1988), whereas NA is responsible for cleaving sialic acid moieties allowing for virion release during replication (Seto et al., 1966). The HA protein is the dominant viral antigen targeted by antibodies (Temoltzin-Palacios and Thomas, 1994) and consists of an immunodominant, highly mutable head domain and a more conserved stalk domain (Wilson et al., 1981). Frequent alteration of HA and NA proteins, both by seasonal mutation (antigenic drift) and via more drastic genetic reassortment (antigenic shift) (Medina and García-Sastre, 2011), generates a highly evasive virus that is extremely difficult to obtain permanent immunity against. Therefore, continued efforts are focused

on gaining a deeper understanding of the humoral response to influenza infection and vaccination.

The best measure available to prevent the significant morbidity and mortality associated with seasonal influenza outbreaks and pandemics, is annual vaccination. In 2015-16, 145 million people received the seasonal influenza vaccine (Santibanez et al., 2015), and the efficacy was around 48% (CDC, 2017). This falls within the upper end of the 19-60% range measured over the last ten years (CDC, 2017) but there is still room for improvement.

The most common seasonal influenza vaccine contains three to four inactivated influenza viruses and is administered intramuscularly. Two of the viruses are from the influenza type A group and the other one or two are from type B. Seven days after vaccination a large expansion of short-lived ASCs, called plasmablasts, can be detected in the peripheral blood (Wrammert et al., 2008). These cells are predominantly antigen specific and of the IgG1 isotype (Wrammert et al., 2008); however, antigen specific IgA plasmablasts are also induced (Mei et al., 2009). It is currently unclear how the vaccine induced IgA plasmablasts relate to the IgG plasmablasts, but further interrogation of this question could help improve vaccine design. The high frequency of BCR mutations within influenza-induced plasmablasts (Wrammert et al., 2008, 2011) indicates prior immune experience where affinity maturation within a germinal center occurred. However, the appearance of these plasmablasts before the peak of germinal center responses (Smith et al., 1996) suggests cellular differentiation through direct activation of pre-existing influenza specific memory B cells (Neu et al., 2016).

Understanding the origin of the responding ASCs is important for vaccine design as this reveals the significance of pre-existing memory in directing a future influenza response (Andrews et al., 2015). Taking this into account, some current vaccine trials aim to shape the influenza memory pool. The goal is to increase the prevalence of broadly neutralizing antibodies over the more common strain specific antibodies (Neu et al., 2016). Unfortunately, recent evidence suggests that continued exposure to antigenically similar strains over time results in increasing the specificity of the response by targeting the highly mutable head of the HA protein (Andrews et al., 2015). To combat this, some current vaccine trials are using a series of immunizations featuring recombinant HA proteins with genetically distinct head domains and conserved stalk domains (Krammer, 2015).

In addition to improving the epitopes targeted, other objectives to improve the protective efficacy of viral immunizations are being explored. One such objective is to target protective antibody responses to mucosal sites where they can block viral entry (Chen and Cerutti, 2010; Ichinohe et al., 2007; Neutra and Kozlowski, 2006; van Riet et al., 2012; Rose, 2014; Wang et al., 2010). For some infections, like HIV, preventing initial infection is the most feasible way to control disease because once an individual is infected, the virus mutates so rapidly that therapeutic antibodies cannot keep up. The prevalence of IgA producing cells in mucosal sites, and the capacity for IgA antibodies to prevent pathogen entry at these sites, makes them an ideal population to target with these vaccines. Unfortunately, our current intramuscular vaccines have been carefully tailored to mount a highly protective peripheral blood IgG1 antibody response. This suggests our current vaccine-induced response is ideal for viral clearance after infection

but has almost no impact on initial entry into the host. If vaccines could be tailored to induce IgA antibodies at mucosal viral entry points it may prevent initial host infection.

1.3 Transcriptional studies to understand vaccine-induced immune responses

The impact of vaccination for human health has fueled substantial research on understanding the immune response elicited by vaccines. A long-standing approach for this characterization relies on performing transcriptome (Chaussabel et al., 2010; Mortazavi et al., 2008; Sultan et al., 2008), or repertoire (Boyd et al., 2009; Ge et al., 2010; Glanville et al., 2009; Ravn et al., 2010; Reddy et al., 2010; Weinstein et al., 2009) studies on total peripheral blood mononuclear cells (PBMCs) or sub-populations that have been identified and isolated by surface marker expression and fluorescence activated cell sorting (FACS) (Fulwyler, 1965; Gray et al., 1975). These studies are responsible for providing significant insight into the immunological subsets responding during a vaccine challenge (Sarkander et al., 2016), as well as our understanding of both the heterogeneity and clonal relatedness of B cell responses against infectious diseases (Georgiou et al., 2014). Additional research has applied these approaches to infections (Liao et al., 2013; Ramilo et al., 2007) allowing for a comparative analysis against vaccines. These approaches have been instrumental in exposing important immune mechanisms to target for vaccine development and therapeutics (Georgiou et al., 2014).

In the last decade vaccine research has turned towards systems biology approaches to glean insight into the systemic immune response occurring after immunization, this is called systems vaccinology (Pulendran et al., 2010). Specifically,

systems biology assumes that no aspect of a biological system can be studied in isolation, as all observations are due to many contributing, intermingled, components (Chuang et al., 2010). Therefore, this field aims to collect broad systemic information and use computational analysis to interpret the biological significance of the complex web of interactions. Systems vaccinology has provided information on the immune mechanisms induced by vaccines, biological correlates of protection and overall signatures of immunogenicity, which are all helping to improve methods for predicting clinical vaccine effectiveness (Pulendran et al., 2010). Early studies identified markers of vaccine efficacy that were expected, such as a strong correlation between neutralizing antibody titers and plasmablast numbers (Nakaya et al., 2011; Querec et al., 2009). Later studies exposed more novel findings, such as the involvement of the innate immune system in live attenuated vaccine responses (Nakaya et al., 2011), and the disparate mechanisms upstream of a protective antibody response for different vaccines (Li et al., 2014; Obermoser et al., 2013). For example, the seasonal influenza vaccine was shown to induce an early interferon signature hours after immunization, whereas the commonly administered pneumococcal vaccine generated a more generic inflammatory response (Obermoser et al., 2013). Further comparative analysis of five different vaccines identified extreme diversity in the immunological mechanisms induced, although all vaccines studied elicited some level of a downstream plasmablast mediated antibody response (Li et al., 2014). In general these studies are critical for vaccine development, as they expose the many different facets of the immune system that need to be considered with each unique pathogen.

Although clearly beneficial, systems biology approaches do have their limitations. The most significant limitation of these studies is that transcriptional profiling of mixed cellular populations is incredibly challenging to interpret, as you cannot differentiate between alterations in cellular frequency and changes in transcript abundance (Chaussabel et al., 2010; Nakaya and Pulendran, 2012). The consequences of this can be minimized when specific bulk populations are analyzed, but at the cost of reduced systemic resolution. In general, studies of the immune system, which is composed of a great diversity of cell types, cannot rely exclusively on these broad approaches, as bulk studies are undermined by the complexity of a heterogeneous immune response.

This heterogeneity is magnified when studying adaptive immune cells due to the unique cellular identity ascribed by the somatically-rearranged antigen receptor that each cell expresses (Murphy and Weaver, 2016). This receptor sequence is only shared with the clonal progeny of the original B or T cell. Identification of clonally related B cells is further complicated because, unlike T cells, their clonal progeny can have significant differences in the BCR sequence due to the point mutations introduced through affinity maturation. Most studies of B or T cells therefore rely on single cell approaches (Georgiou et al., 2014). One of the most successful approaches to study the repertoire and functional capacity of these adaptive receptors has been to use PCR-based receptor cloning from single cells followed by the study of purified antigen receptors in vitro (Smith et al., 2009; Wang et al., 2012; Wrammert et al., 2008). The successful application of single cell approaches to adaptive immune cell characterization has primed the field of immunology for the rapid assimilation of single cell RNA-sequencing

(scRNA-seq) technologies, which allow for a broader description of the gene expression profiles that dictate cellular function.

1.4 Single cell RNA-sequencing [1]

In the past decade advances in computing and sequencing technologies have ushered in a new era of discovery in immunology. In particular, scRNA-seq has enabled an unprecedented view of gene expression in single cells (Jaitin et al., 2014; Tang et al., 2009). For cells of the immune system, at least two computational and bioinformatic challenges exist for effectively harnessing single cell genomics data. The first challenge lies in visualizing single cell gene expression data in a biologically meaningful way while remaining robust to the high levels of noise that is present in single cell data. The second challenge lies in exploiting single cell sequencing data to infer immune repertoire, and thus information on cell specificity and function.

As mentioned above, the use of single cell based technologies in immunology has a rich history. However, scRNA-seq is revolutionary with its ability to offer an entirely unbiased strategy for identification and characterization of distinct cellular populations through quantification of all transcripts expressed in a single cell. Recent immunological studies utilizing scRNA-seq have identified novel subpopulations (Björklund et al., 2016; Gaublomme et al., 2015; Meredith et al., 2015; Shalek et al., 2013), and elaborated cellular differentiation pathways (Olsson et al., 2016; Zhou et al.,

[1] This section is largely reproduced from Neu, K.E, Tang, Q., Wilson, P.C. and Khan, A.A. (2017). Single-Cell Genomics: Approaches and Utility in Immunology. Trends in Immunology *38*, 140-149. License number 4092061175096 was granted by Elsevier.

2016). For instance, scRNA-seq was used to uncover the developmental states in myelopoiesis and infer the underlying regulatory genes (Olsson et al., 2016). scRNA-seq has also revealed previously unappreciated cellular functions (Mahata et al., 2014; Shalek et al., 2014), such as up-regulation of the steroid synthesis pathway in Th2 cells following helminth infection, resulting in suppression of Th cell proliferation and B cell immunoglobulin class recombination (Mahata et al., 2014). Additionally, scRNA-seq has helped to unveil the complexities of tumor microenvironments (Powell et al., 2012), exposing a connection between transcriptional heterogeneity within tumors and prognostic outcome (Patel et al., 2014; Tirosh et al., 2016).

The applications of scRNA-seq are extensive and proper execution of data analysis is essential. Significant effort over the past few years has resulted in an initial framework to process and analyze scRNA-seq data. However, many important challenges remain unsolved for the effective use of scRNA-seq data. Here some of the outstanding technological limitations specific to scRNA-seq and unique computational considerations in data analysis for immunology research will be discussed (Figure 1.2).

1.4.1 scRNA-seq data processing

A common theme among scRNA-seq approaches is use of oligo recognition of the mRNA polyA tails for cDNA generation, followed by amplification and subsequent sequencing (Saliba et al., 2014). Once high throughput reads from a well-designed scRNA-seq experiment are obtained, there is a plethora of analysis packages available as well as many quality control measures (Bacher and Kendziorski, 2016; Grün and van Oudenaarden, 2015). Broadly, quality control should be performed on the reads themselves and across the cells in an effort to prune low quality data prior to analysis.

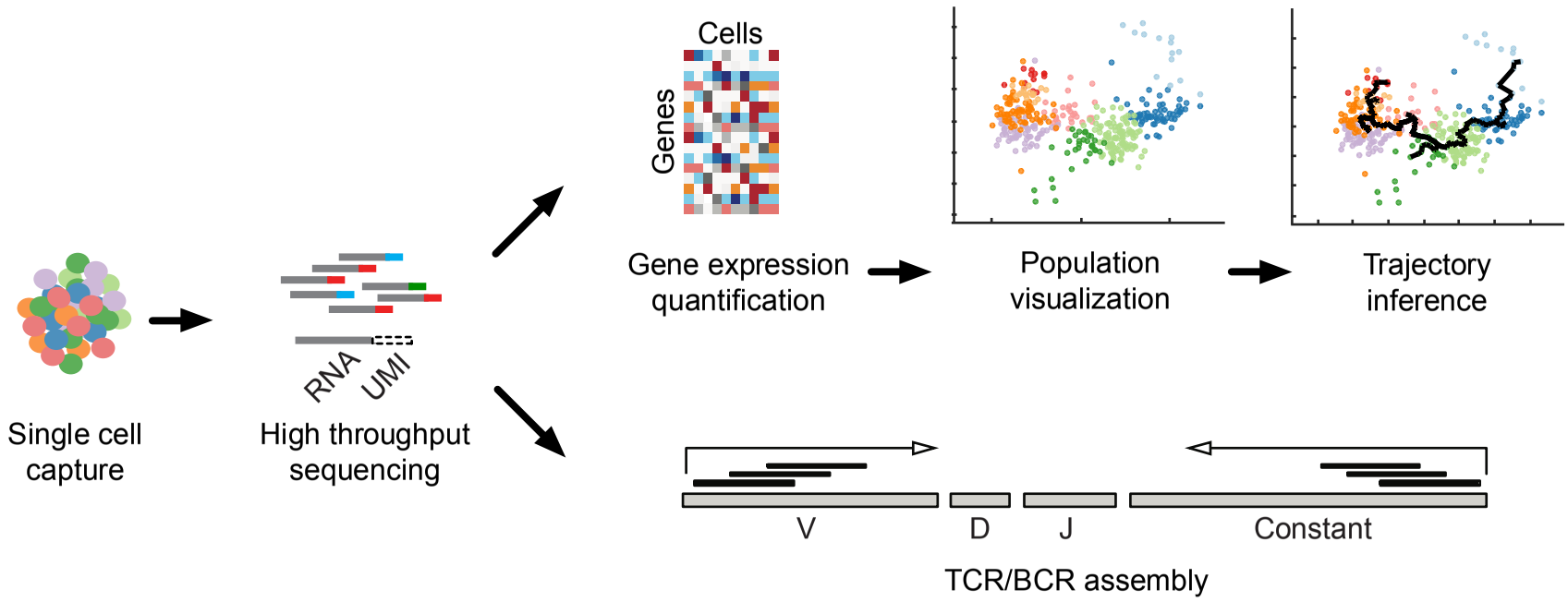


Figure 1.2 scRNA-seq analysis outline. The standard workflow of scRNA-seq studies begins with isolation of single cells, followed by library generation and high-throughput RNA sequencing. Multiple computational and bioinformatics tools exist for quantifying gene expression. In general, these tools report the amount of reads that are associated with a specific gene, typically normalizing for gene length and library size. Before gene expression quantification, it is important to perform quality control analyses, including filtering for read quality and eliminating cells with overall low library size. After quantification, batch effect, dropout effect, and amplification bias should also be considered in normalizing scRNA-seq gene expression data (see Figure 1.4.2). Ultimately, scRNA-seq data can be subjected to dimensionality reduction algorithms to reveal different cell subpopulations, potentially infer their developmental trajectory during a dynamic process, and, if applicable, use sequencing data to identify their adaptive receptor sequence to assess questions of clonality and specificity. BCR, B-cell receptor; TCR, T-cell receptor; UMI, unique molecular identifier.

Therefore, an important element of engaging in scRNA-seq studies is the use of specialized computational and bioinformatic tools for quality control and data processing.

This section reviews typical methods for gene expression quantification and considers special challenges related to scRNA-seq data processing, including: batch effects, dropout effects, and amplification bias (Figure 1.3). Accurate gene expression quantification is a key prerequisite for downstream analysis of scRNA-seq data. The importance of data processing can be emphasized by invoking the longstanding cautionary statement “garbage-in garbage-out”, as unaddressed technical biases and effects can drastically undermine scRNA-seq studies.

1.4.1.1 Gene expression quantification

A number of the tools available for bulk RNA-seq processing have been successfully adopted by scRNA-seq. For gene expression quantification some examples include; STAR, RSEM, the Tuxedo Suite**,** and Kallisto (Bray et al., 2016a; Dueck et al., 2015; Langmead and Salzberg, 2012; Li and Dewey, 2011; Ntranos et al., 2016). These approaches map sequencing reads to a reference genome, a transcriptome index or perform *de novo* assembly without a reference genome to allow for expression quantification. The gene expression is typically reported as raw counts, RPKM/FPKM or TPM, which are different but related measurements. Counts are simply a count of the reads mapping to a gene, whereas the other units have undergone an additional normalization step on a per-cell basis to account for library depth as well as gene transcript length (Li et al., 2010; Mortazavi et al., 2008). Roughly, these measurements represent a value reflecting the units of a gene transcript in a cell. There is ambiguity in

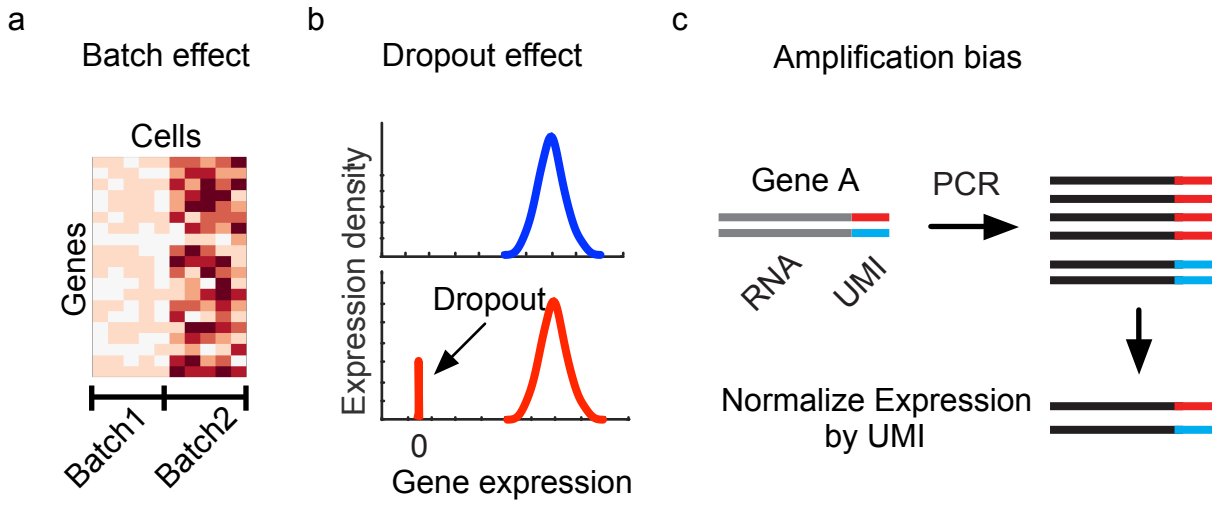


Figure 1.3 scRNA-seq Specific Challenges. scRNA-seq is by nature a fragile and sensitive technique. Experimental challenges inherent to this approach can confound data analysis and should be accounted for. (**a**) First, technical batch effects can cause the data to artificially cluster by batch, overshadowing the biological variability. (**b**) Second, dropout effect causes an enrichment of non-biological zero values that results in a bimodal expression profile, which can impair subpopulation analysis. (**c**) Finally, certain genes may experience preferential gene amplification, introducing a bias that is difficult to adjust for. The use of unique molecular identifier (UMI) can help account for amplification bias and allow for more accurate quantification of starting RNA levels.

the field regarding which measurement to use, and additional subsequent forms of normalization are recommended for accurate comparison across samples (Bacher and Kendziorski, 2016; Zyprych-Walczak et al., 2015).

1.4.1.2 Batch effect

Among the most challenging aspects of harnessing data from scRNA-seq is the processing of multiple datasets that have been produced across different batches (Hicks et al., 2015; Tung et al., 2016). These batches originate from sample number limitations inherent in most scRNA-seq protocols. Batch effects can confound experimental variables of interest and have long been appreciated for their dangers in gene expression analysis (Johnson et al., 2007; Zyprych-Walczak et al., 2015). The primary danger in ignoring batch effects can be illustrated by erroneous situations where different cell types in a single batch demonstrate greater correlation than identical cell types from different batches (Gilad and Mizrahi-Man, 2015; Hicks et al., 2015; Vandenbon et al., 2016) (Fig. 1.3). Here, a statistical model can be used to directly subtract bias from known batch assignments (Johnson et al., 2007; Leek and Storey, 2007). In particular, ComBat uses parametric and non-parametric empirical Bayes frameworks to describe the expression of genes through an additive combination of batch effect variables (Johnson et al., 2007). Despite technological advances increasing the number of samples in each batch, as seen with Drop-seq (Fan et al., 2015; Klein et al., 2015; Macosko et al., 2015), multiple batches are still commonly used to assess cellular heterogeneity or to reproduce earlier findings; thus batch correction maintains its relevance (Macosko et al., 2015).

One way to assess the impact of batch effects on the experimental data is to use

principal component analysis (PCA) to examine how variance along the principal dimensions segregates the single cell data by known batch variables (Leek et al., 2010). If single cell data appears to segregate with the batch in which the cells were processed, more effort may be needed to remove batch effects, including using surrogate variables for identifying and estimating unknown sources of variation and adjusting the data to remove these effects (Leek and Storey, 2007). This adjusted dataset can be used for downstream analysis, where biological differences should no longer be confounded by experimental batch. However, the trade-off is that greater efforts to remove batch effects may impose homogeneity in the data and impair or reduce true signals associated with differentially expressed genes.

1.4.1.3 Dropout effect

Due to the relatively low levels of RNA present in single cells, several biases manifest in scRNA-seq data. One problem that bears importance is a dropout effect. Here, identical cells processed through scRNA-seq may yield different gene expression profiles with absent expression for some genes due to technical limitations of the amplification process (Kharchenko et al., 2014; Pierson and Yau, 2015; Prabhakaran et al., 2016; Shalek et al., 2013). For example, the PCR may have been unable to amplify the gene, potentially due to competition between genes preventing comprehensive amplification. This is detected by an inflation of zero values which results in bimodal gene expression profiles (Shalek et al., 2013) (Fig. 1.3). Here, removal of cells through a quality control process, where a baseline number of genes needs to be present, may be a reasonable method to proceed (Gaublomme et al., 2015; Grün and van Oudenaarden, 2015). An emerging trend in computational analyses of scRNA-seq data is to specifically account

for dropout effects by imputing missing values using probabilistic models, such as in ZIFA (Pierson and Yau, 2015) and BISCUIT (Prabhakaran et al., 2016). The BISCUIT algorithm probabilistically models the entire data using cell-specific scaling factors to impute missing values. Filling in these missing values can help reveal biologically relevant trends in data, previously obscured by dropout-missing values. Although quality control and imputation algorithms help to address this bias, it is likely technological modifications to the sequencing chemistry may be needed to fully address dropout effects.

1.4.1.4 Amplification bias

Concomitant with dropout effect is another concern known as amplification bias. Here, specific genes may (or may not) preferentially amplify due to the non-random priming of sequences. This problem is pronounced for scRNA-seq due to the higher number PCR cycles typically involved in amplifying RNA, and generating cDNA, material from a single cell. One mechanism to try and account for amplification bias in scRNA-seq data is to use unique molecular identifiers (UMI) or barcodes prior to amplification (Kivioja et al., 2012). The principle with this approach is that unique barcodes will randomly ligate to each molecule of starting RNA, allowing a better estimate of original transcript numbers after amplification and controlling for gene specific bias (Figure 1.3). However, this is susceptible to other unforeseen biases such as ligation bias, and recent data suggests UMI's remain susceptible to biological and technical variation (Tung et al., 2016). A complementary method to address sample-specific amplification bias involves the use of External RNA Control Consortium (ERCC) spike-in controls, where the RNA sample is "spiked" with some known quantity of RNA (Jiang et al., 2011). Then, after

quantification, gene expression profiles can be normalized against spike-in levels. Unfortunately, spiked-in RNA adds another layer of complexity to scRNA-seq protocols that is expensive in time and resources and remains susceptible to the same amplification and batch effects as all other genes. It is important to note that UMI and spike-ins, whether used separately or together, are still unable to resolve dropout effects since cell specific technical variations, such as RNA extraction efficiency, will occur before barcoding or introduction of spike-ins.

There remains a critical need in developing effective computational methods to normalize scRNA-seq data so that biological variability in gene expression can be quantified in an accurate manner.

1.4.2 Visualization of scRNA-seq data

One of the most compelling applications of single cell genomics to immunology resides in characterizing the population structure of single cells. Visualization of scRNA-seq data can help to identify rare and intermediate subpopulations that are overlooked with bulk RNA-seq data. The goal of visualization algorithms is to project high-dimensional data into a low dimensional space, thus resolving cellular groups based on their transcriptional similarity without any pre-determined markers dictating their identity (Figure 1.2). In this section we review two standard dimensionality reduction algorithms used to visualize scRNA-seq data and highlight their applications to immunology.

PCA is a linear dimensionality reduction algorithm with long standing popularity for high-dimensional data (Jolliffe et al., 2002). PCA takes an input of correlations between cells based on gene expression data, and identifies principal components corresponding to linear combinations of genes, which cumulatively capture the

variability of the total dataset. When the data is projected against the first few components, which account for the largest amount of variation, distinct populations can be visually and biologically interpreted. This approach has led to improved characterization of haematopoietic stem cells (Zhou et al., 2016) and Th17 heterogeneity (Gaublomme et al., 2015), among many others. A convenient aspect of PCA analysis is that information on the contribution of each gene to the component is maintained allowing for further biological interrogation of the clusters (Ringnér, 2008). One example exploited this powerful feature when detecting cellular heterogeneity within the autoimmune Th17 compartment (Gaublomme et al., 2015). Previously identified cell subset transcriptional information was projected into their PCA space to resolve PC1's connection to differentiation between effector to memory T cells and PC2's connection to cellular origin (Gaublomme et al., 2015).

A comparable non-linear algorithm is called t-SNE, t-Distributed stochastic neighbor embedding (Maaten and Hinton, 2008). Unlike PCA, which seeks to capture variance in data, t-SNE seeks to explicitly preserve the local structure of the original data. t-SNE constructs a probability distribution to describe the dataset such that pairs of similar cells are assigned a high probability, while dissimilar pairs are assigned with a much smaller probability. Thus, cells that are similar in the high-dimensional space will cluster together (due to high probability) in low-dimensional space. This ability to explicitly maintain clustering of similar cells is an advantage of t-SNE over direct linear transformation such as PCA. This approach is very effective with scRNA-seq data, and has been used for characterizing subpopulations in innate lymphoid cells (Björklund et al., 2016), the tumor-immune microenvironment (Cann et al., 2012; Patel et al., 2014;

Powell et al., 2012; Tirosh et al., 2016), as well as exposing Aire-dependent transcriptional program stochasticity (Meredith et al., 2015). Specifically, when looking at innate lymphoid cell populations the use of t-SNE was able to resolve a third transcriptionally distinct population that had been indistinguishable with PCA (Björklund et al., 2016).

Despite the successes of t-SNE in effectively visualizing scRNA-seq data, there are two algorithmic drawbacks that are critical for immunologists to note. First, due to the stochastic nature of t-SNE, the same dataset may produce different visualizations in different runs (Maaten and Hinton, 2008). Thus, it is important to run t-SNE on the same dataset multiple times in order to obtain a strong intuition of the population structure. Second, while t-SNE maintains clusters by placing cells that are similar in the original space together in the low-dimensional space, cells that are dissimilar in the original space may not necessarily be placed proportionally apart in low-dimensional space (Maaten and Hinton, 2008).

Overall, dropout effects remain a challenge for dimensionality reduction algorithms such as PCA and t-SNE. In particular, PCA may provide misleading results in the single cell setting, where the first component may seek to explain variability in data due to dropout effects (Hicks et al., 2015). Algorithms that impute the missing values have shown great promise with dimensionality reduction algorithms, by significantly improving the separation of distinct subpopulations; thus facilitating differential expression and other downstream analyses (Pierson and Yau, 2015; Prabhakaran et al., 2016).

1.4.3 Immune repertoire

Characterizing the diverse repertoire of an ongoing immune response has proven quintessential to our understanding of human health. This was first tackled at the single cell level with cell fusion hybridomas in the 80s and eventually progressed to PCR based receptor chain cloning and recombinant protein expression. These tools exposed key components of the basic biology behind T and B cells, some examples include; VDJ rearrangement in T cells (Kavaler et al., 1984), SHM (Weigert et al., 1970), and the identification of memory B cells (Klein et al., 1998) and analysis of single B cell specificities (Wardemann et al., 2003). These approaches have also had significant therapeutic impact with the identification of broadly neutralizing antibodies (Wilson and Andrews, 2012) and in some current cancer immunotherapy approaches (Littman, 2015).

Now, scRNA-seq has exposed new opportunities to analyze the immune repertoire and its impact on immune cell function, simultaneously allowing both for capture of quantitative gene expression as well as repertoire sequence information for both chains (α/β, light/heavy) of the receptor (Figure 1.2). The assembly of full-length BCR or TCR sequences from scRNA-seq is a non-trivial problem. On the one hand, reference-based assembly methods that rely on an initial alignment of reads to a genome are vulnerable to the somatic rearrangements and mutations. On the other hand, de novo assembly methods typically require solving a noisy and complex genome-scale assembly task that makes targeted assembly error-prone. Here we review the available methods for assembling TCR or BCR sequences from scRNA-seq data.

TraCeR (Stubbington et al., 2016) was the first tool for TCR reconstruction from scRNA-seq, followed quickly by VDJPuzzle (Eltahla et al., 2016). BASIC (discussed in detail in section 2.2) is currently the only algorithm designed specifically for B cells and the only approach that performs its own assembly of the V(D)J receptor sequence (Canzar et al., 2016). All three approaches follow the same two main steps: a filter step and an assembly step. The filter step acquires the RNA-seq reads that align to all possible gene segments used in V(D)J recombination, or align to a comprehensive database of artificial recombinomes utilizing these genes. The assembly step is crucial, as the receptor sequence cannot be identified by alignment alone due to the large amount of junctional diversity introduced during receptor rearrangement. For this, TraCeR and VDJPuzzle both utilize a popular *de novo* transcriptome assembly software package, Trinity, to construct the TCR (Grabherr et al., 2011). Alternatively, BASIC uses an identified conserved sequence anchor in both the 5' (V gene) and 3' (Constant region) end of the receptor to perform semi-*de novo* BCR assembly, by extending these anchors in either direction with varying k-mers of overlapping sequence. In most cases the sequence from the constant region and variable region for each chain will eventually overlap, resulting in a single transcript being reported for each chain. However, if the sequences do not meet, both a constant and variable contig will be reported for that chain. The receptor sequences generated by all three of these algorithms were confirmed with PCR-based receptor amplification followed by Sanger sequencing, and demonstrate high accuracy with murine splenic $CD4^+$ T cells, human peripheral blood antigen specific $CD8^+$ T cells, and human peripheral blood plasmablasts (Smith et al., 2009; Wang et al., 2012).

This novel ability to link repertoire and transcriptome information at the single cell level exposes many new opportunities for investigation, such as links between transcriptomes and clonality, receptor specificity, receptor affinity, receptor gene usage, complementarity determining region (CDR) loop variability, BCR mutation load, and BCR isotype. Already, this innovative combination showed that the T cell clonotypes responding to a murine Salmonella infection come from diverse T cell compartments and it identified a previously unappreciated repertoire diversity within HCV specific T cells before and after *ex vivo* cell culture (Eltahla et al., 2016; Stubbington et al., 2016). Prior interrogation of both questions was impossible, and it will be very exciting to see what else this powerful new combination brings in the future.

1.4.4 Concluding remarks on scRNA-seq

ScRNA-seq has paved the way for new avenues of research, with unprecedented power. Unfortunately, due to the relative youth of the field, only a limited number of tools have been developed to address the complexity of high-dimensional scRNA-seq data sets. Further, there exists a lack of standardization for analyses in the field. Regardless of the difficulties associated with data analysis, the field of immunology has embraced this new tool with fervor, resulting in outstanding recent discoveries (Bendall et al., 2014; Buettner et al., 2015; Cann et al., 2012; Gaublomme et al., 2015; Mahata et al., 2014; Patel et al., 2014; Powell et al., 2012; Shalek et al., 2013, 2014; Tirosh et al., 2016). Needless to say, there remain many unanswered questions that eagerly await the application of scRNA-seq. While not yet routine, recent advances in microfluidics are allowing significantly larger numbers of cells to be barcoded and processed for scRNA-seq (Fan et al., 2015; Jaitin et al., 2014; Klein et al., 2015; Macosko et al.,

2015). As the magnitude of data obtained from these experiments continues to grow, and the popularity of the approach spreads, scRNA-seq will likely become a standard method for studying immunological cellular heterogeneity.

1.5 Aims and significance

Single-cell RNA-sequencing is ushering in a new era for research, where previously unappreciated heterogeneity will be unearthed. Single cell approaches have experienced great popularity within immunology and the application of these approaches to human B cells has contributed significantly to our understanding of the humoral immune system. However, knowledge of the adaptive antigen receptor of a cell only imparts a partial identity. In order to merge this pre-existing characterization with total transcriptome profiling and advance our understanding of vaccine immune responses, I set out to accomplish the following set of specific aims for my dissertation:

Aim 1: Generation of an integrative framework for simultaneous single cell RNA-seq, B cell receptor sequencing and monoclonal antibody generation.

Aim 2: Application of this approach to human peripheral blood plasmablasts following influenza vaccination.

Here, I present the development of our integrative framework Spec-seq and introduce BASIC, an algorithm we developed that identifies full-length BCR sequences from within scRNA-seq datasets. The application of Spec-seq to human plasmablasts following influenza immunization identified immunoglobulin genes as the major drivers of heterogeneity and exposed previously unappreciated transcriptional changes associated with BCR antigen specificity. This approach also facilitated the first

transcriptional analysis of clonally-related human B cells, the results of which may suggest original antigen exposure permanently alters the transcriptome. These findings, and the potential applications of this framework, have the capacity to significantly advance our understanding of the adaptive immune system.

2. Spec-seq

2.1 Introduction

As discussed in section 1.3, significant advances in the field of immunology have come from performing transcriptome and repertoire studies on large populations of cells. These studies can provide critical information on the dynamics of the ongoing immune response and the clonal relationships between cells. The advent of scRNA-seq technologies have recently introduced new avenues for exploration (see section 1.4). This level of resolution is particularly valuable when studying the cells of the adaptive immune system, because each cell has it's own unique functional identity that is signified by its antigen receptor. However, the ability to simultaneously interrogate these features at the individual cell level has remained largely unappreciated.

To address this gap in the B cell field, we developed Spec-seq, an integrative framework that allows for simultaneous monoclonal antibody extraction and transcriptional profiling from the same cell. Not only does this technique facilitate joint repertoire and transcriptome profiling from a single cell, it also allows for downstream functional characterization of the BCR expressed by that cell. This provides a completely unique ability to link gene expression profiles to the antigen specificity and therapeutic potential of an individual cell. This protocol is a substantial contribution to the field of immunology and constitutes a significant proportion of my thesis accomplishments (Fig. 2.1).

Importantly, the joint transcriptome – repertoire dataset obtained via Spec-seq is innovative even without further receptor functional characterization, since it provides the first dataset where B cell transcriptome can be directly linked to receptor sequence.

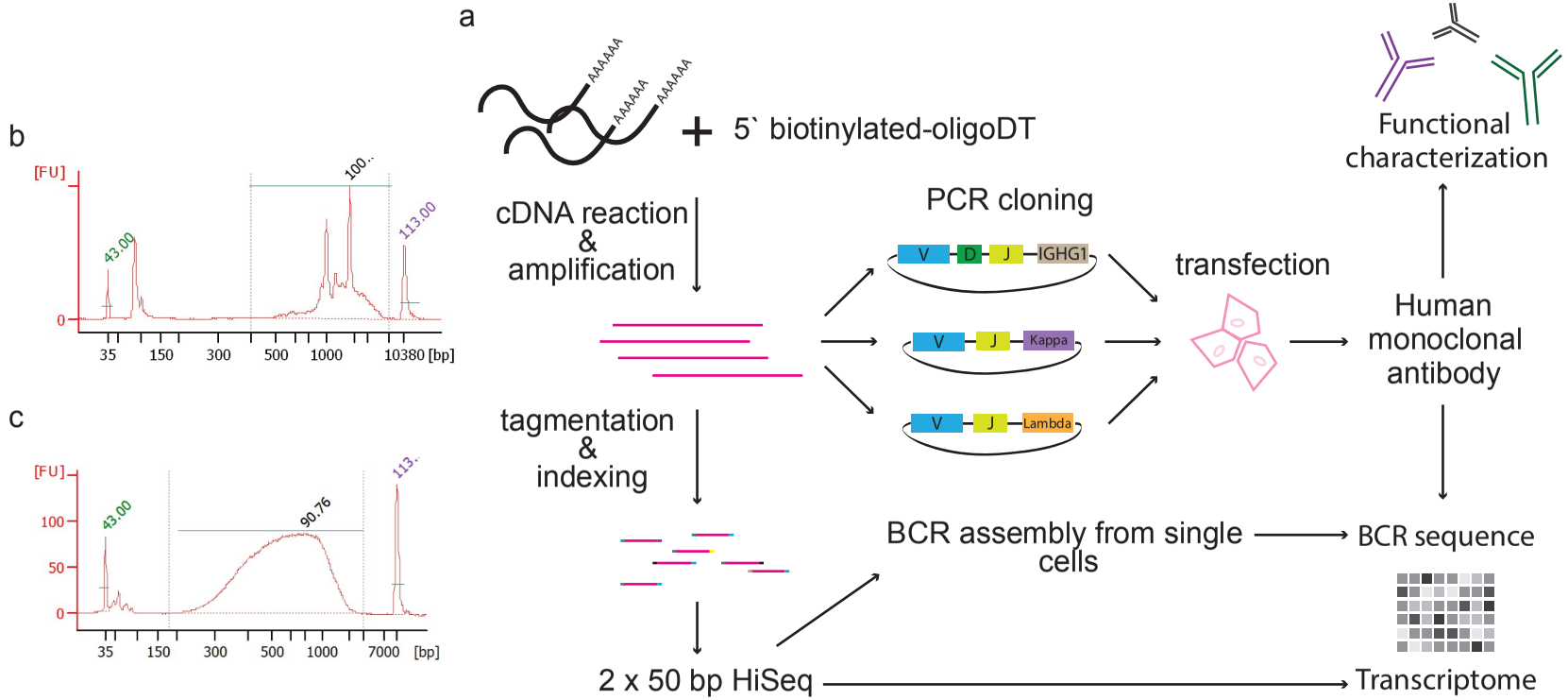


Figure 2.1 Spec-seq method details. (**a**) Spec-seq allows for simultaneous analysis of B cell receptor sequence, receptor binding properties and transcriptional profile from a single cell. This is accomplished by performing PCRs, to amplify the BCR, on an aliquot of full-length amplified cDNA from each cell. The cDNA is generated via reverse transcription from total cell mRNA captured by a poly-adenenlated oligomer, and amplified with 22 additional PCR cycles. The amplified V(D)J sequence is cloned into an expression vector, matched heavy and light chains are transfected into HEK293A cells and a fully human monoclonal antibody is generated. Indexed libraries are generated from 0.5 ng of the remaining cDNA prior to next generation sequencing (**b**) Representative Agilent bioanalyzer high sensitivity chip results from the chosen 22 cDNA amplification cycle number. (**c**) Representative Agilent chip results after DNA tagmentation. All plasmablast traces contain the two sharp amplicon peaks suggesting extremely high expression of certain transcripts, the smaller of which we attribute to the BCR itself.

Although scRNA-seq has begun to be applied to different B cell populations, the high level of genetic complexity of the BCR makes it impossible to identify from within the transcriptome data without the use of complex algorithms. Therefore our dataset provided a unique platform to develop the BASIC algorithm, which can perform complete, highly accurate, BCR assembly from within scRNA-seq datasets.

2.2 Development and overview of Spec-seq

In this section I will briefly outline the Spec-seq protocol and explain the validation steps involved in the optimization for human peripheral blood plasmablasts. Additional details can be found in Chapter 5 - Experimental procedures.

The cell population of interest was sorted directly into 96-well plates containing 4 µl of the Smart-seq2 cell lysis solution (Picelli et al., 2014). Subsequently all polyadenylated messenger RNA (mRNA) is amplified from the cell lysate and cDNA is generated. This process relies on a reverse transcription enzyme that introduces known un-templated nucleotides at the end of each cDNA piece. These nucleotides will immediately anneal to the three riboguanosines in the provided TSO-oligo, allowing the enzyme to continue transcription. This is known as template switching and results in the same oligomer sequence at the 5’ and 3’ end of each cDNA piece, facilitating unbiased downstream amplification.

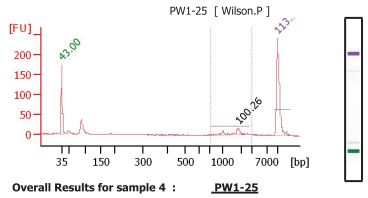
In order to determine the optimal number of pre-amplification cycles for peripheral blood plasmablasts a number of trial experiments were performed. In these experiments single cells, and 100 cell wells, were processed and amplification was assessed on the Agilent high sensitivity DNA chip after 18, 22, 24, 26 and 29 pre-

amplification cycles were performed (Fig. 2.2). 22 pre-amplification cycles was selected as optimal because it offered decent amplification of single cells while remaining within an acceptable range for the 100 cell wells. One factor we ascribed to over amplification was when the primer/oligo peak became significantly larger than the first ladder peak, as was seen with 24, 26 and 29 cycles (Fig. 2.2).

Every sample included for RNA-seq was therefore processed with 22 pre-amplification cycles and with quality assessment on the Agilent high sensitivity DNA chip. This functions as our first cell-quality filtering step. Although the results obtained from our single cells were extremely variable we were able to set clear requirements for inclusion in downstream library generation (Fig. 2.3). We required the average cDNA fragment size to be between 1400 and 2000 base pairs. Additionally, if the cDNA concentration was significantly below 100 pg/µl the sample was also eliminated, because the necessary 0.5ng could not be achieved within the 5µl volume limitation for library generation. We observed improvement in the bioanalyzer results when we switched from the SuperScriptIII (Fig. 2.3a) to the Primescript enzyme (Fig. 2.3b).

Within all plasmablast traces analyzed for this project there was a consistent double peak within the amplified cDNA. The absence of these peaks in similarly processed memory B cells (Fig. 2.4a) has led us to propose they are due to two very highly expressed plasmablast specific transcripts. Specifically, we believe the smaller of these two peaks is likely to be the BCR itself, which is known to be expressed at exceptionally high levels within antibody secreting cells and would be decreased in memory B cells.

SC - 18



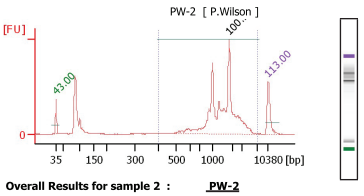
Overall Results for sample 4 : PW1-25

Number of peaks found:	1
Noise:	0.5
Corr. Area 1:	79.0

Region table for sample 4 : PW1-25

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
716	2,959	79.0	26	1,624	33.3	26.07	27.3	

SC - 22



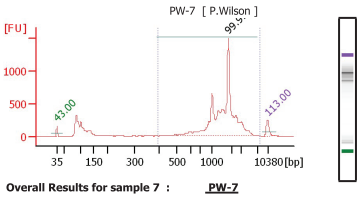
Overall Results for sample 2 : PW-2

Number of peaks found:	1
Noise:	0.5
Corr. Area 1:	2,371.4

Region table for sample 2 : PW-2

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
399	6,898	2,371.4	72	1,660	55.7	720.74	892.5	

SC - 24



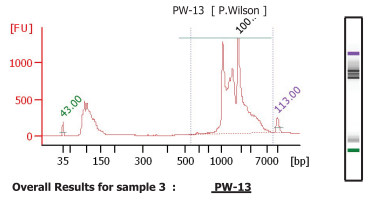
Overall Results for sample 7 : PW-7

Number of peaks found:	1
Noise:	0.7
Corr. Area 1:	6,117.8

Region table for sample 7 : PW-7

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
395	8,043	6,117.8	74	1,747	55.0	2,062.68	2,434.9	

SC - 26



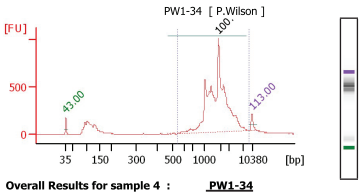
Overall Results for sample 3 : PW-13

Number of peaks found:	1
Noise:	0.8
Corr. Area 1:	8,164.5

Region table for sample 3 : PW-13

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
538	9,047	8,164.5	68	2,066	62.8	3,246.05	3,004.7	

SC - 29



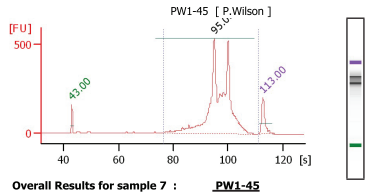
Overall Results for sample 4 : PW1-34

Number of peaks found:	1
Noise:	2.7
Corr. Area 1:	6,082.8

Region table for sample 4 : PW1-34

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
539	9,128	6,082.8	81	2,192	63.6	2,929.41	2,688.9	

HC - 18



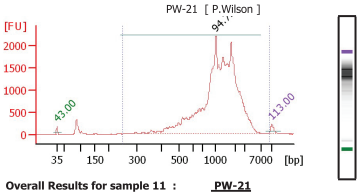
Overall Results for sample 7 : PW1-45

Number of peaks found:	1
Noise:	0.8
Corr. Area 1:	3,512.4

Region table for sample 7 : PW1-45

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
393	8,460	3,512.4	93	1,442	52.3	1,257.91	1,728.3	

HC - 22



Overall Results for sample 11 : PW-21

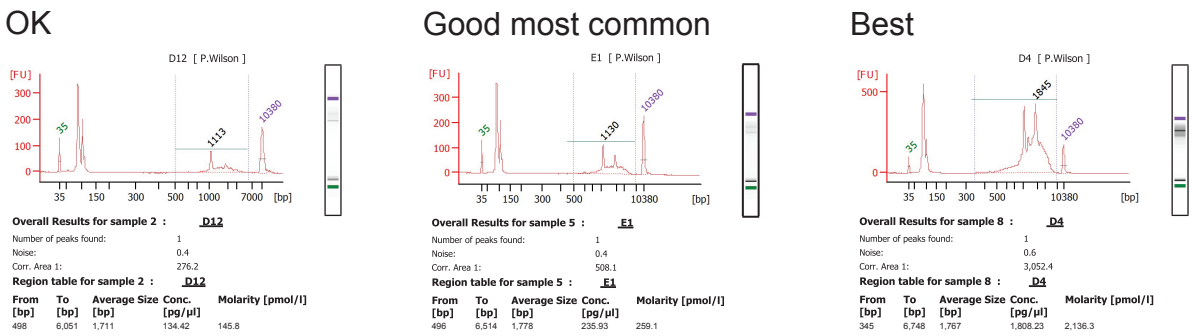
Number of peaks found:	1
Noise:	0.6
Corr. Area 1:	25,756.4

Region table for sample 11 : PW-21

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
250	9,490	25,756.4	96	1,453	65.2	10,015.36	16,076.6	

Figure 2.2 Pre-amplification cycle number titration. Agilent bioanalyzer traces after different PCR-preamplification cycle numbers; 18, 22, 24, 26, 29. Representative figures are either single cell (SC) or one hundred cells (HC). Each plot has the same general compoents. The ladder peaks are visible around 35bp/42s and then again around 10380bp/115s. The next peak, which has a lot of variaibility between samples and cycle numbers is the oligomer/primer peak. The total amplified cDNA is the large peak contained mostly between 300-7000bp/80-110s. A unique double peak was observed within the cDNA region for all plasmablast samples processed.

a



b

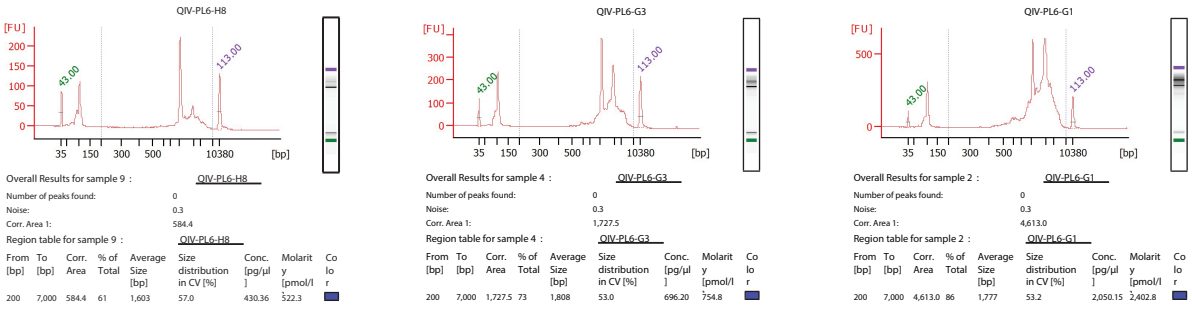
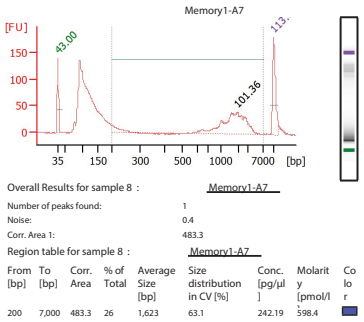


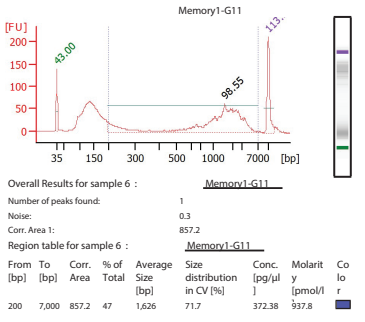
Figure 2.3 Comparison of SuperScriptIII and Primescript enzymes. (**a**) Chip results from SuperScriptIII enzyme. The first and second plots show fairly common amplification peaks achieved with this enzyme, with the third plot representing a more infrequent better quality peak. (**b**) Chip results with the primescript enzyme show better amplification plots, with the second trace being the most common result obtained from cells amplified with this primer.

a

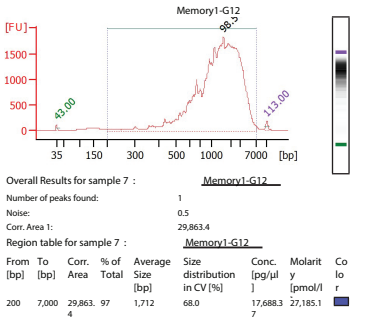
Single cell



Single cell

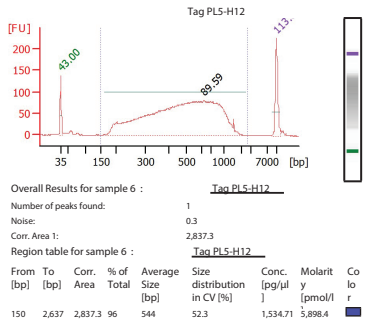


100 cell well

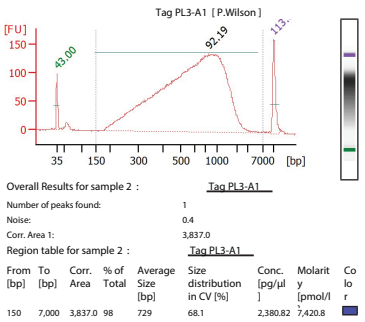


b

OK common



Better common



Best infrequent

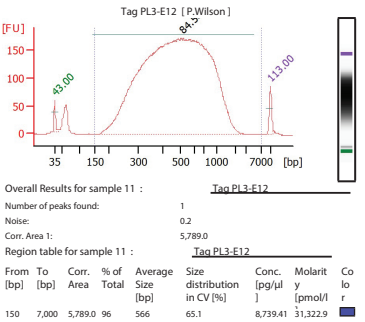


Figure 2.4 Memory and tagmented libraries. (**a**) Peripheral blood memory B cell (CD19+, CD27+, CD38+) Agilent chip results. (**b**) Final plasmablast tagmented library chip results. The first two are common results obtained in this dataset, showing a wide spread of DNA fragment sizes, with the amjority trending towards longer than 500 bp. The third trace is an ideal result, with an even spread of fragment sizes centered on 500 bp.

After confirmation of successful cDNA amplification a small aliquot of cDNA was removed from each sample and used in PCR reactions designed to amplify the heavy, kappa and lambda chains of the BCR (Smith et al., 2009). This step functions as our second cellular quality control step. The requirement for a successfully cloned antibody selects only wells that contain a single high-quality cell, as PCR amplification of the BCR fails on empty wells or wells containing more than one cell. All cells without successfully amplified BCR chains were either excluded from scRNA-seq submission or removed from the final RNA-seq data directly. The Sanger sequencing obtained during the process of amplifying the BCR provides key repertoire information on each cell, including (but not limited to) V(D)J gene usage, mutation levels, CDR3 sequence and receptor isotype. Subsequent cloning and transfection of the identified BCR generates a fully human monoclonal antibody, identical to the one produced by the original cell, except that all heavy chain variable regions are cloned into the IgG1 backbone. The functional characteristics of the antibody can then be assessed.

Cells that passed both of these quality control steps then underwent the remaining steps for library generation. This process was tightly batch controlled, meaning the 96 samples that would be sequenced together were processed together in a single plate, and that plate layouts were carefully planned for optimal interrogation of the question of interest. The product of the tagmentation and indexing reaction is again submitted for Agilent high sensitivity DNA chip analysis. Ideally, all cDNA fragments should now be around 500 bp, with our results ranging from ~200-1500 bp (Fig. 2.4b). Many of the samples included in our early experiments had low concentrations and

more uneven fragment size at this step; however, downstream sequencing appears very similar and of high enough quality for the questions at hand.

2.3 BASIC - BCR assembly from single cells[1]

The ability of Speq-seq to independently identify the true BCR sequence provided a unique dataset for the development of a tool that extracts this information directly from the tagmented scRNA-seq reads. The algorithm was tailored throughout design by comparing its assembled algorithms against the known BCR sequence. Without the novel data set generated with Speq-seq, the creation of BASIC would have been impossible.

As mentioned in section 1.4.3, the assembly of full-length BCR sequences from scRNA-seq is a complex problem due to the extreme sequence diversity introduced by V(D)J recombination (Tonegawa, 1983), SHM (Rajewsky et al., 1987; Weigert et al., 1970) and CSR (Lawton et al., 1975). During V(D)J recombination additional genetic variability is realized through junctional diversity or the inexact joining of gene segments. The receptor heavy chain is unique in its inclusion of the 'D' gene segment, which generates increased genetic diversity within the antigen-recognizing domain. This locus therefore presents a novel computational challenge. On the one hand, reference-based assembly methods that rely on an initial alignment of reads to a genome are

[1] This section is largely reproduced from Canzar*, S., Neu*, K.E, Tang, Q., Wilson, P.C. and Khan, A.A. (2016). BASIC: BCR assembly from single cells. *Bioinformatics*, 33(3), 425-427. *The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as Joint First Authors. No license was required for re-use of this material.

vulnerable to the somatic rearrangements and mutations present in BCR sequences. On the other hand, *de novo* assembly methods typically require solving a noisy and complex genome-scale assembly task that makes targeted BCR assembly error-prone. Thus, the lack of efficient methods for assembling BCR sequences from scRNA-seq data is a major roadblock in advancing B-cell biology.

Here, we present our novel semi-*de novo* assembly method to determine the full-length sequence of the BCR in single B cells from scRNA-seq data, called BASIC (BCR assembly from single cells). To demonstrate the utility of our method, we subjected nearly 200 single human B cells to scRNA-seq and assembled full-length heavy and light chains. We then experimentally validated these results by using single-cell RT-PCR and Sanger sequencing, and demonstrate overall high accuracy.

2.3.1 BASIC algorithm

We developed BASIC to determine the full-length sequence of the heavy and the light chains in the BCR of a single B cell from scRNA-seq. Briefly, BASIC performs semi-*de novo* assembly in two stages (Fig. 2.5b): Stage 1, BASIC uses known constant and variable regions to identify anchor sequences; Stage 2, BASIC uses these anchors to guide the *de novo* assembly of the BCR.

Stage 1: The BASIC software uses a pre-compiled database of known variable (IGHV, IGKV, IGLV) and constant (IGHC, IGKC, IGLC) region sequences in human from IMGT (http://www.imgt.org). The database was indexed using Bowtie2 into four distinct files that corresponded to different components (IGHV; variable heavy chain, IGHC; constant heavy chain, IGK/LV; variable light chain, IGK/LC; constant light chain). Second, BASIC calls Bowtie2 (Langmead and Salzberg, 2012) in order to align scRNA-seq reads from a

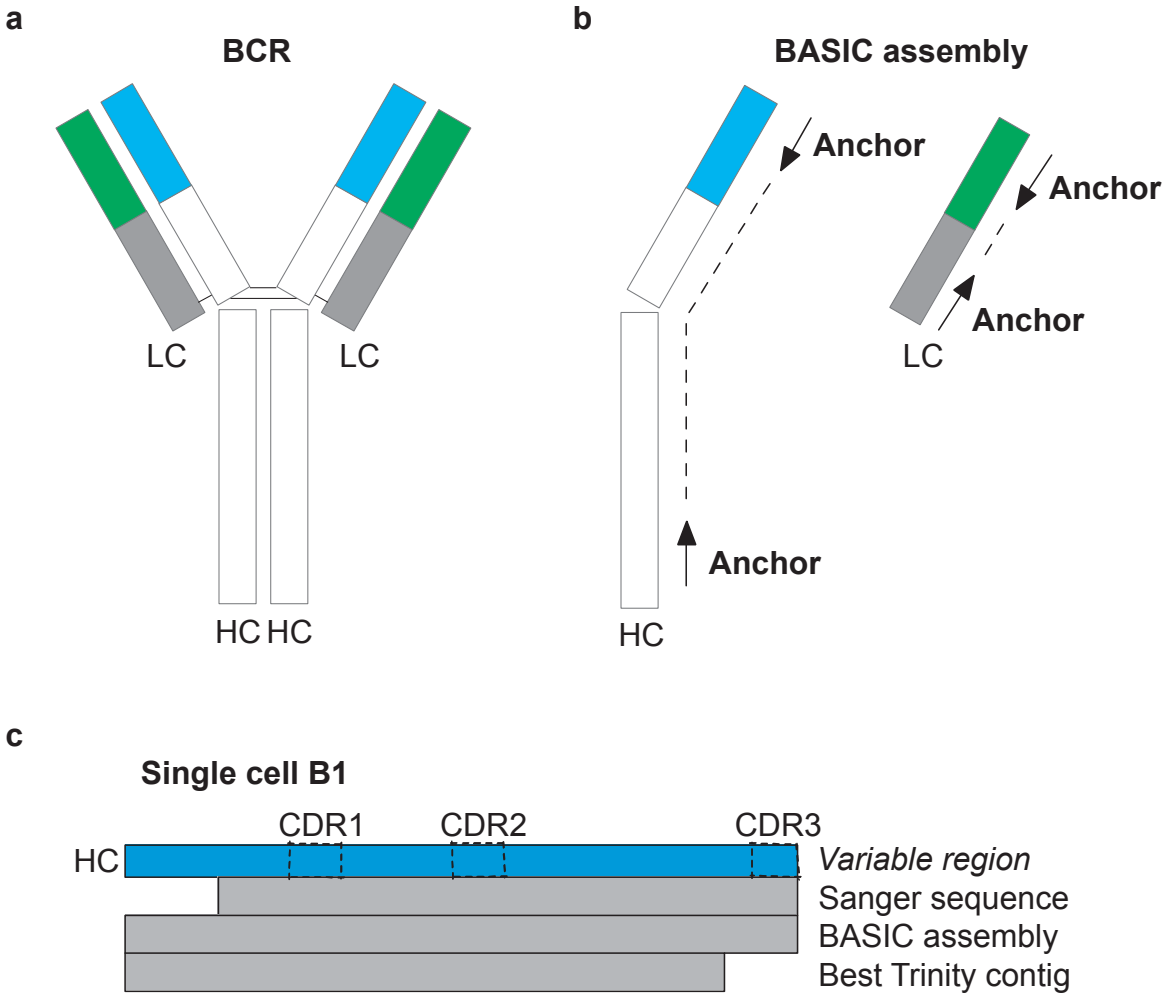


Figure 2.5 BASIC assembly schematic and contig comparison. (a) The BCR is a complicated protein dimer composed of two identical heavy chains (HC) and two identical light chains (LC). The complementarity determining regions (CDRs) are at the top of both arms of the antibody and are the parts of the variaible regions that participate in the binding of antigens. (b) BASIC identifies conserved sequence anchors that are stitched together to assemble the HC and LC. (c) Illustration of the HC variaible region sequence for single-cell PW1-B1 along with the Sanger sequence, the BASIC assembled sequence and the best contig reported by Trinity.

single B cell to each of the four component index files. Third, for each component, BASIC identifies a short sequence window containing the highest number of aligned reads. These four sequences served as anchors to guide the assembly stage. As expected for anchors in the variable regions, sequences typically mapped to the non-complementarity determining regions (non-CDR). Lastly, we filtered all remaining low-complexity reads for which at least half of the read sequence was composed of a single nucleotide.

Stage 2: Next, BASIC performs *de novo* assembly to stitch together the anchor sequences in the heavy and light chains. We assume a sequence may overlap either with the forward sequence or the reverse complement sequence of another read. Two reads overlap if the prefix of one sequence equals the suffix of the other sequence (or vice versa). Briefly, BASIC extends each anchor iteratively in the 3' direction (one read at a time) until there is either no overlapping read or a repeat is found. Then, each anchor is extended in the 5’ direction in the same way. For each chain, BASIC reports a single sequence if the extended sequence from the variable region anchor is equal to the extended sequence from the constant region anchor. However, if the two sequences did not match, BASIC reports both an extended variable region contig and an extended constant region contig, separately. Two contigs will result, for example, when there is not enough sequencing read coverage spanning the entire BCR.

To efficiently extend the anchor sequences we developed a greedy minimum entropy approach. In each iteration, starting with the anchor sequence, we identified all overlapping reads for a given direction. This can be seen as generating a star graph with one internal node (current sequence) and edges to all possible reads overlapping

the current sequence. In contrast to other assembly graphs that use a fixed k-mer size, we conceptually explore a graph structure that connects nodes by an edge if the corresponding read sequences overlap by any length k. In that way, we are able to handle uneven sequencing coverage of the BCR transcript, which is common in scRNA-seq.

The running time of BASIC mostly depends on the length of the sequencing reads and the number of reads. BASIC is implemented in Python and uses multi-core processing to assemble the heavy and light chains simultaneously.

2.3.2 Experimental validation of BASIC assembled sequences

We obtained plasmablasts by fluorescence-activated cell sorting from the peripheral blood of seven human donors. These samples were subjected to scRNA-seq using an adapted version of the SmartSeq2 protocol (Picelli et al., 2014). Some of these adaptations included testing different pre-amplification PCR cycles (Column B, Table 2.1), and the removal of a small aliquot of amplified full-length cDNA to validate BASIC predictions using nested PCR-based BCR amplification and Sanger sequencing (Smith et al., 2009; Wrammert et al., 2008). The nested PCR utilized a cocktail of primers for all possible V and J genes. ScRNA-seq libraries were then prepared by tagmentation, and single-end (17 samples) or paired-end (174 samples) 50 base-pair sequencing was performed on an Illumina HiSeq2500 machine. In total, our data set contains 191 single cells, which resulted in 382 V(D)J sequences (Table 2.1).

In 37 situations our initial nested PCR was unable to amplify the BCR using a cocktail of primers, potentially due to competition between primers preventing amplification or a mutation in the primer-binding site (Column D, Table 2.1). This

resulted in a unique opportunity to apply and test BASIC, where we then selected specific gene primers based on the prediction of the BCR from scRNA-seq data. Consequently, we performed a series of secondary PCRs by using specific primers to amplify the BCR based solely on BASIC and successfully recovered all 37 sequences (Column AA, Table 2.1).

2.3.2.1 BASIC is highly accurate in predicting BCR sequences

BASIC successfully identified V(D)J gene usage as obtained by Sanger sequencing in both heavy and light chains in 370/382 (~97%) of the sequences. Importantly, BASIC was able to correctly identify the V(D)J sequence across a range of somatic alterations and sequencing depths (Columns M, Y, AJ, Table 2.1).

BASIC failed to correctly assemble 12/382 (or ~3%) of the immunoglobulin sequences. It stands to reason that sample quality and, in turn, sequencing coverage of the BCR is a major determinant for accurate assembly. While BASIC does not explicitly use pairing information, the additional sequences generated from paired-end sequencing likely provides for greater coverage of the BCR and implicitly facilitates the *de novo* assembly stage in BASIC. Consistent with this, 5/12 of the incorrect predictions were obtained from single-end sequencing data. Taken together, BASIC correctly assembles full-length BCR sequences across a range of sequencing and experimental conditions including somatic mutations, sequencing depths, and pre-amplification PCR cycles.

2.3.2.2 BASIC harnesses anchors to guide accurate BCR assembly

We sought to determine the importance of using anchors to guide assembly in our semi-*de novo* approach. We assembled all 191 scRNA-seq samples using Trinity, a state-of-

the-art, purely *de novo* approach. Because a limitation in using a purely *de novo*-based method is identifying BCR related sequences among the hundreds of thousands of contigs assembled, we assumed we had an oracle to tell us the single best contig by aligning all of Trinity's contigs to the true reference BCR sequence using BLAST.

We evaluated V(D)J gene usage in the *de novo* assembled contigs returned by our oracle. We found 31 instances where our oracle failed to return a correct immunoglobulin sequence in contrast to BASIC (Fig. 2.5c; Column AK, Table 2.1). At the same time, there were only 4/12 instances where our oracle successfully returned a correct sequence when BASIC could not. In sum, our oracle failed to return a correctly assembled sequence among Trinity's contigs for 42/382 (or 11%) of the immunoglobulin sequences (Table 2.2). Taken together, our analysis demonstrates the importance of our guided, anchor-based approach.

2.3.3 Discussion

The BCR is generated through an immunological process known as affinity maturation and is ultimately secreted in soluble form as antibodies. During this antibody affinity maturation process, somatic rearrangements and mutations are introduced into the BCR to improve antigen recognition. We have introduced a novel algorithm, BASIC, which enables investigators to assemble BCR sequences from scRNA-seq data and study B-cell repertoire at single cell resolution. The optimal design of this algorithm depended on prior knowledge of the BCR sequence, as well as experimental validation in cases where the BCR sequence was not obtained with the PCR approach. BASIC demonstrates overall high accuracy and in particular the superiority of an anchor-guided assembly strategy over a purely *de novo* based approach. For the first time, BASIC can

Table 2.2
Summary of errors for BCR assembley from Trinity:

Categories of Trinity Errors	# errors
Incomplete BCR assembly	21
No BCR assembly	10
BCR assembled, incorrect genes	6
BCR assembled, incorrect genes (same as BASIC)	5
Failure to assemble BCR	31
BCR assembled but incorrect	11
Total errors	42

help couple gene expression information from scRNA-seq with immune repertoire, and facilitate studies between B-cell receptor features, clonality, differentiation and transcriptional programming at single cell resolution.

2.3.4 Contributions

S.C and K.E.N are joint first authors. S.C and A.A.K designed the BASIC algorithm. K.E.N performed experiments and assessed assembly results. P.C.W oversaw and funded lab experiments. A.A.K oversaw and directed this project. S.C, K.E.N & A.A.K wrote the manuscript.

2.4 Conclusion

Spec-seq opens new doors for research in the adaptive immune system where single cell approaches are already routine. Although our work here focused on B cells, this protocol could easily be modified to work on other cells as long as the receptor primers are available for PCR amplification. Fortunately, even if the primers are not available, tools like BASIC (Canzar et al., 2016, Elthala et al., 2016 and Stubbington et al., 2016) are quickly being developed and can provide repertoire information without the lab intensive PCR approach. In these cases downstream receptor synthesis could be used to facilitate functional characterization.

Spec-seq is an integrative framework constructed from multiple pre-existing tools. These tools are incredibly powerful on their own (Picelli et al., 2014 & Smith et al., 2009), but their influence has been upgraded through this merger. Significant troubleshooting was required to optimize the scRNA-seq element of this approach to work within our lab and for our population of interest. Fortunately, this approach is now

being successfully applied to other questions at the University of Chicago (see Chapter 4).

In particular I expect this joint repertoire-transcriptome information will be highly beneficial for personalized medicine research and therapeutics. This revolutionary direction has gained significant clinical popularity for cancer and autoimmunity, where the identification of clonal expansions are key for diagnosis, and where understanding pathologic, or tumor infiltrating, cellular identity is essential for selecting proper treatment options. Spec-seq could advance this field.

On a smaller scale this approach will now facilitate interrogation of additional B cell populations within human and murine systems, as well as of more complex B cell events like affinity maturation or class switch recombination.

3. Receptor antigen reactivity and clonal origin alter the gene expression profiles of human antibody secreting cells.

Vaccines are the most effective public-health tool for combating infectious diseases such as influenza. The humoral immune system is key to vaccine efficacy yet, our understanding of how antibody specificities relate to individual B cell function remains limited due to the complexity of polyclonal antibody responses. Single cell RNA-sequencing has recently enabled simultaneous transcriptome and B or T cell receptor sequence analysis (Canzar et al., 2016; Eltahla et al., 2016; Stubbington et al., 2016), but the connection between antigen-binding specificities and transcriptional profiles remains equivocal. Here we present the first application of the Spec-seq framework, which allows for simultaneous mAb characterization and transcriptional profiling from the same single cell. We applied this to human plasmablasts after influenza vaccination in order to characterize transcriptional differences dictated by B cell receptor (BCR) isotype and vaccine reactivity. Our results suggest that plasmablasts from the same clonal progenitor have more similar transcriptomes than unrelated cells. This enhanced similarity extends beyond the BCR itself suggesting that original B cell activation events are imprinted on the transcriptome and that this is preserved in downstream clonal progeny. The direct comparison of IgG and IgA vaccine positive plasmablasts identified clustering caused by BCR constant domain transcripts, but detected no further transcriptional specialization associated with the class switch recombination event. Alternatively, the comparison of vaccine positive and vaccine negative plasmablasts identified transcriptional subpopulations reflective of BCR vaccine recognition. Differentially expressed genes between these populations identified glycotransferase enzymes previously credited with altering the glycosylation of antibodies, thereby

altering their immunogenicity, as well as differences in cell cycle activity. These data suggest IgA vaccine negative cells are transcriptionally distinct from traditional terminally differentiated antigen-induced peripheral blood plasmablasts.

3.1 Introduction

Plasmablasts are the main antibody secreting cells following recent antigen challenge. Steady state plasmablasts are present at a low level in the peripheral blood of healthy humans and undergo significant expansion seven days post immunization. The vaccine-induced population is predominantly antigen reactive and has been extensively utilized as a way to characterize the dynamics of an ongoing B cell immune response (Smith et al., 2009; Wrammert et al., 2008). This population is also a common source for therapeutic and prophylactic human mAb therapies (Wilson and Andrews, 2012). The vaccine-induced plasmablasts typically express the IgG receptor isotype (Wrammert et al., 2008) in contrast to the steady-state plasmablasts which mostly express the IgA receptor isotype (Mei et al., 2009). Although CSR alters the functional capacity of antibodies, it remains unknown if the process alters the transcriptional profile of the plasmablast, beyond changes to the BCR transcript itself.

The steady-state IgA secreting plasmablast population is maintained in the peripheral blood during an ongoing vaccine response (Mei et al., 2009). Interestingly, how these cells are related to and how they differ from the antigen-induced cells remains unclear. The peripheral blood steady-state plasmablasts are hypothesized to be recirculating from mucosal niches (Mei et al., 2009), where they secrete antibodies that function to maintain homeostasis of the microbiota (Pabst, 2012). The recent

discovery of shared IgA antibody sequences in both the serum and the gut, further supports the connection between steady-state circulating IgA secreting cells and mucosal localized cells (Iversen et al., 2017). This possible shared origin, combined with the enhanced neutralization capacity of IgA antibodies, makes them an ideal target population for vaccines aimed to induce mucosal protection (Ichinohe et al., 2007; Neutra and Kozlowski, 2006; van Riet et al., 2012; Rose, 2014; Sundararajan et al., 2015; Wang et al., 2010), making the exploration of IgA plasmablasts (steady-state versus antigen-reactive) of great interest.

Overall, gaining a deeper understanding of any transcriptional specialization associated with B cells producing antibodies with disparate features could advance the design of vaccines, therapeutic antibodies and perhaps regulation of the microbiome.

Spec-seq is a framework to allow for the simultaneous exploration of the immune transcriptome, receptor repertoire and receptor functional characteristics at the single cell level (Fig. 3.1a & Fig. 2.1). As a first application of this we address the long-standing questions of how BCR isotype or specificity impacts the B cell transcriptome, and if these features direct functional specialization outside of the BCR itself. Specifically, we examine how IgA vaccine positive (IgAVP) peripheral blood plasmablasts compare to IgG vaccine positive (IgGVP) and IgA vaccine negative (IgAVN) populations (Fig. 3.1b).

3.2 Results

IgA and IgG peripheral blood plasmablasts were single-cell sorted from six patients, seven days after influenza vaccination (Fig. 3.1c, d & Table 3.1). As expected with an

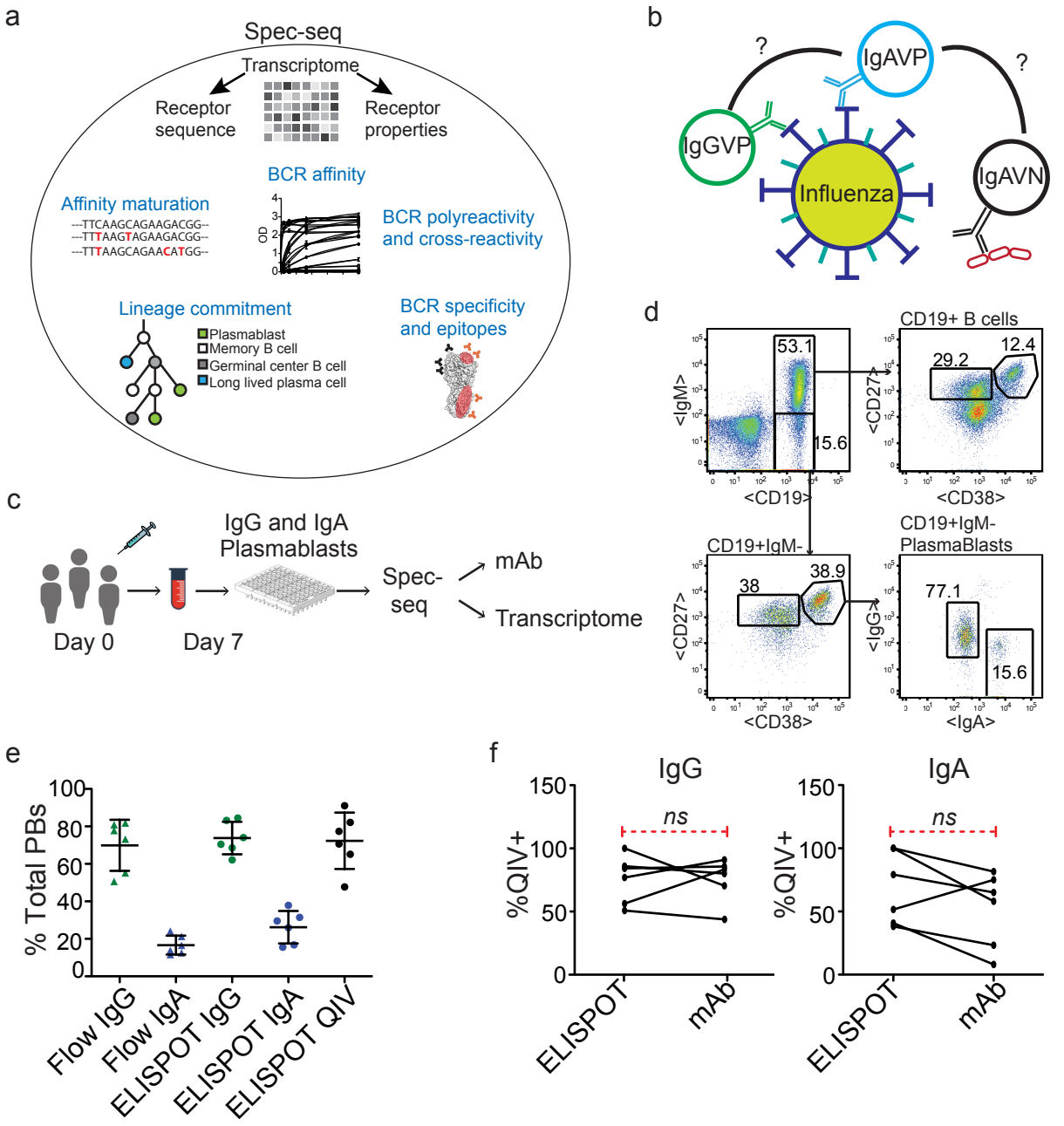


Figure 3.1 Spec-seq overview and application. (**a**) Potential applications of Spec-seq. (**b**) The IgA vaccine positive (IgAVP) plasmablasts will be compared against IgG vaccine positive (IgGVP) and IgA vaccine negative (IgAVN) plasmabalsts. (**c**) Experimental outline. (**d**) Sorting strategy, IgM- plasmablasts were identified as CD19+CD27++CD38++ and either IgG+ or IgA+. (**e**) The frequency of total plasmablasts of each isotype as detected during cell sorting and through anti-Ig ELISPOT. The ELISPOT data also reports the frequency of vaccine specific plasmabalsts. Line indicates frequency mean. (**f**) The per-patient frequency of vaccine specificity by ELISPOT and mAb, separated by isotype.

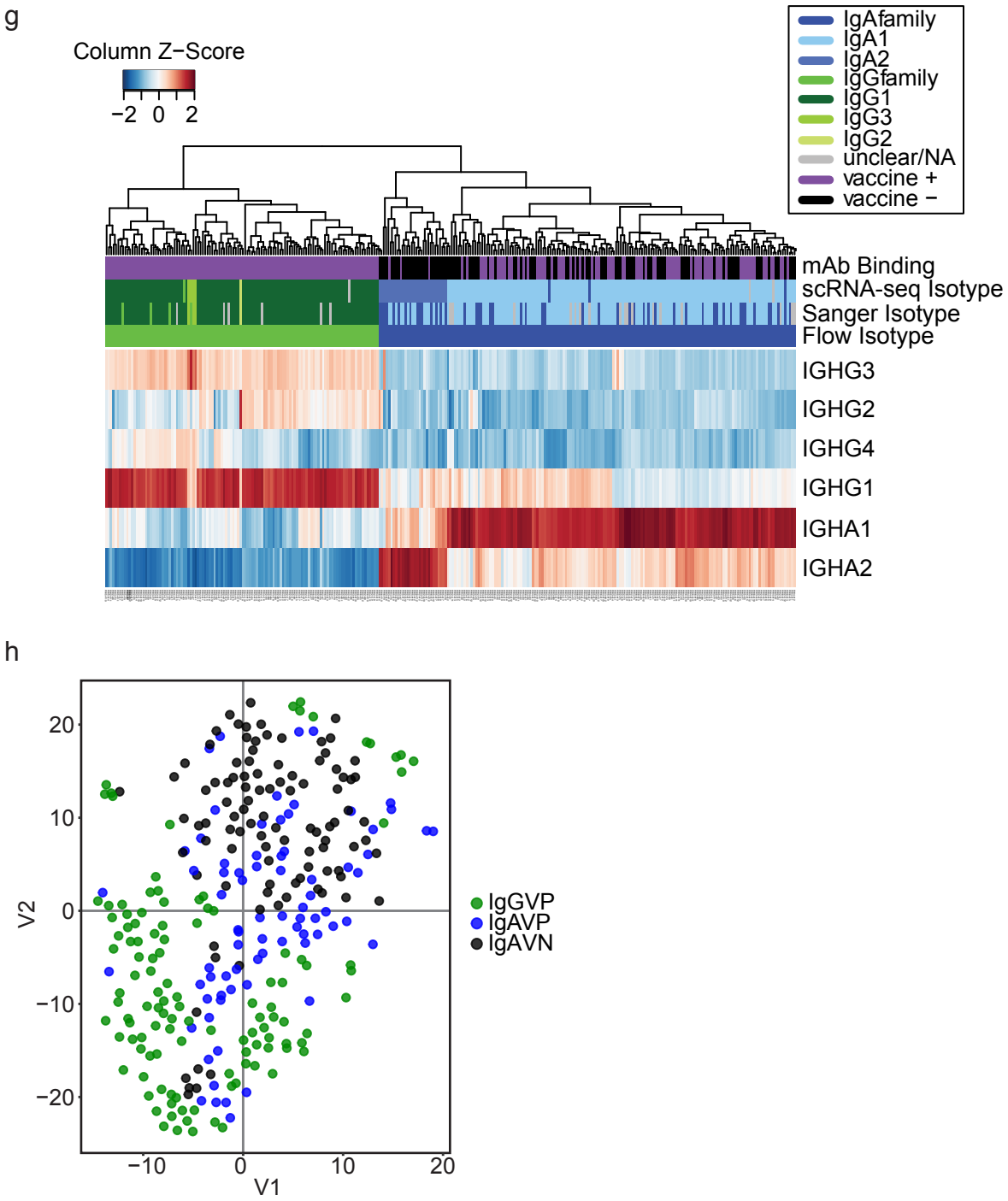


Figure 3.1 Continued (**g**) Heatmap of all 295 single cells, clustered by their relative expression of the four IgG subtype genes and the two IgA subtype genes. Annotation bars across the top show the mAb vaccine binding and BCR isotype. (**h**) tSNE projection of the entire transcriptome of all 295 cells colored by group ID.

intramuscular vaccine, there was a wide range of plasmablast expansion per patient (1.6-26.7%) with the majority of plasmablasts expressing IgG (flow cytometry 51-82%, ELISPOT 62-85%) (Fig. 3.2a, & Fig. 3.1e). The total frequency of vaccine binding plasmablasts per-patient ranged from 48-91% (Fig. 3.1e). Monoclonal antibodies (mAb) were cloned and generated from each plasmablast and screened by ELISA for binding to the administered vaccine (Fig. 3.2b). Although not significant, the mAb vaccine positive frequency tended to be reduced from the frequency detected by ELISPOT (Fig. 3.1f) and we attribute this to the inability of our total vaccine ELISA to sample all possible antigen epitopes derived from the vaccine *in vivo* (Kuraoka, Masayuki, 2016; Wrammert et al., 2008). Sorted plates containing single cells were subjected to deep single cell RNA sequencing (scRNA-seq) (see Chapter 5). We focused our subsequent analyses on 295 plasmablasts that passed quality control (QC) filters with 11895 unique genes detected in at least 5% of cells. We applied the BASIC algorithm (Canzar et al., 2016) to assemble the full-length Ig receptor sequences.

The mAb binding data combined with the receptor sequence information allowed each cell to be binned into one of our three groups of interest: IgAVP, IgGVP and IgAVN (Fig. 3.1g & Table 3.2). The BCR isotype classification was validated on three levels with flow cytometry, Sanger sequencing during BCR cloning and the assembled contig from BASIC. The scRNA-seq data was normalized (see Chapter 5 & Fig. 3.3) and visualized using the t-Stochastic Neighbor Embedding algorithm (tSNE), which projects high dimensional data into a two-dimensional space and groups cells with similar transcriptomes together (Maaten and Hinton, 2008). As expected, after proper normalization and batch correction, clustering is not associated with experimental batch

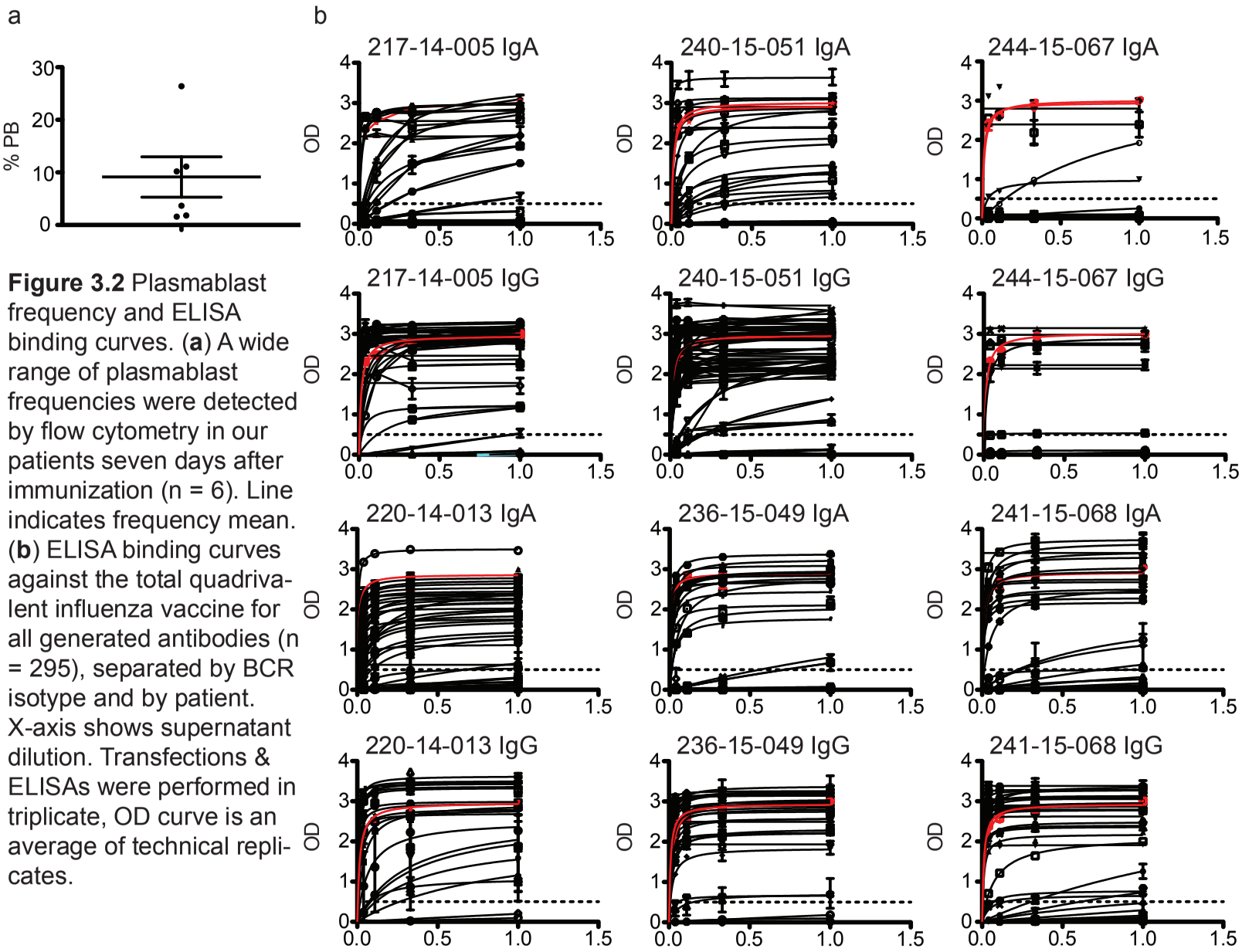


Figure 3.2 Plasmablast frequency and ELISA binding curves. (**a**) A wide range of plasmablast frequencies were detected by flow cytometry in our patients seven days after immunization (n = 6). Line indicates frequency mean. (**b**) ELISA binding curves against the total quadrivalent influenza vaccine for all generated antibodies (n = 295), separated by BCR isotype and by patient. X-axis shows supernatant dilution. Transfections & ELISAs were performed in triplicate, OD curve is an average of technical replicates.

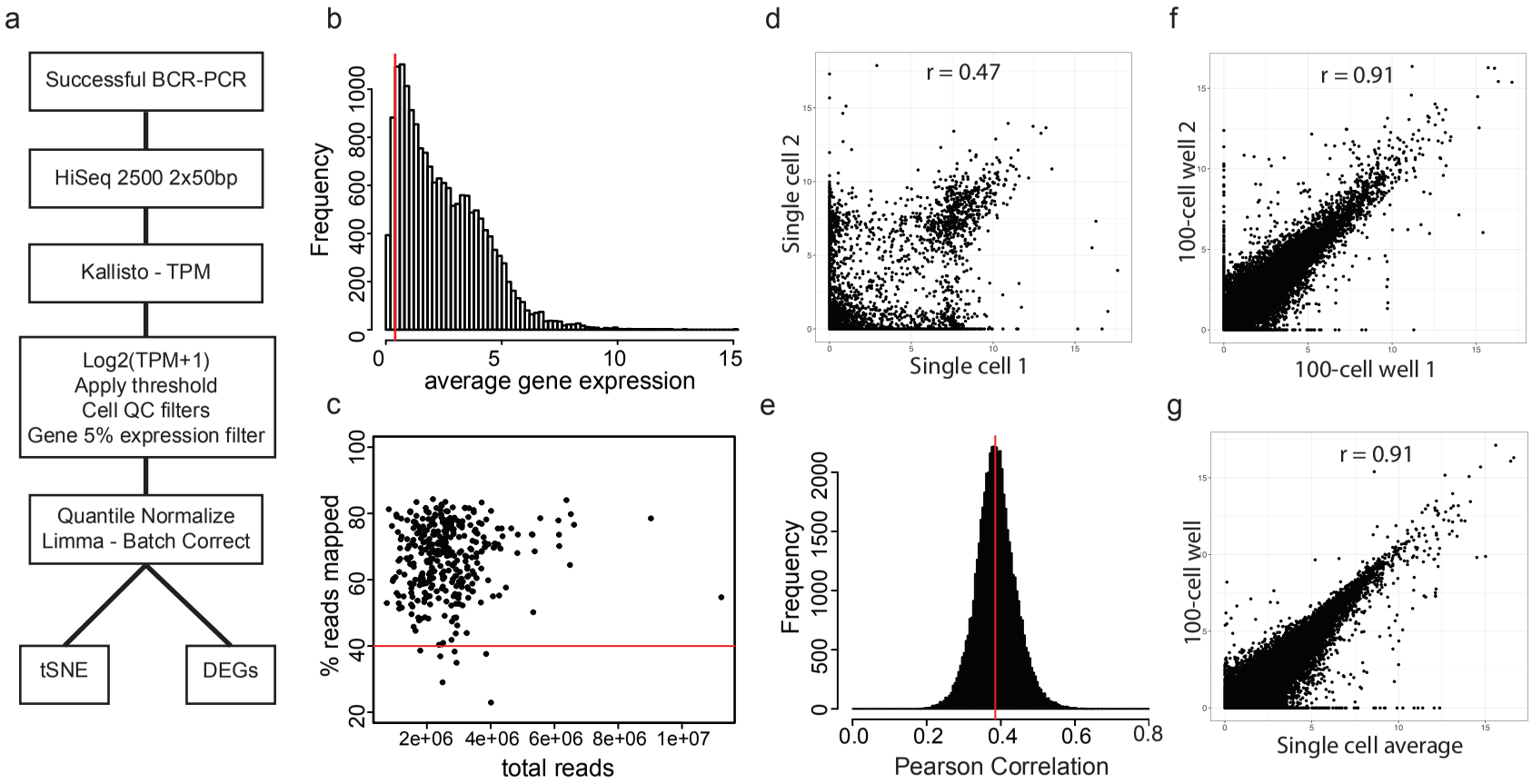


Figure 3.3 Analysis pipeline, filters and quality control.(**a**) Schematic of the overall data analysis pipeline. (**b**) Histogram of the average gene expression, obtained after excluding zeros. Red line indicates the threshold of detection, all values below this threshold were set to zero. (**c**) Any cell with less than 40% transcripts mapping to the human transcriptome were excluded from analysis. (**d**) A representative scatter plot of two single cells. (**e**) A histogram of all single cell pairwise correlations, red line at median correlation. (**f**) The Pearson correlation between two 100-cell wells is 0.91, which is the same as when the average of all single cells is plotted against a 100-cell well (**g**). This suggests that although single cells have varying levels of heterogeneity their composite profile still captures the population profile.

or patient (Fig. 3.4). Instead, a trend was revealed for the IgAVP and IgAVN cells to cluster together and away from the IgGVP population (Fig. 3.1h), suggesting a plasmablast BCR isotype-dependent signature.

Immunoglobulin variable gene repertoire analysis revealed little bias between the three populations, and did not suggest distinct cellular compartments of origin (Fig. 3.5 & Supplementary Tables 3.1-3.2). However, we did identify an increased frequency of IgAVN cells using the IgA subtype 2 constant domain (*IGHA2*), which is more resistant to protease degradation and is therefore more common within mucosally localized antibodies (Spencer and Sollid, 2016). We also identified an increased isoelectric point within the complementary determining three region of the BCR within the IgAVN cells. This is known to contribute to polyreactivity, which was recently exposed as a mechanism utilized by IgA ASC populations for recognition of the microbiome (Bunker et al., 2017). Both of these characteristics support a possible connection between the peripheral blood IgAVN population and the mucosal IgA ASC population. Repertoire analysis also revealed evidence of clonotype expansion and identified 100 of the total 295 plasmablasts as members of 29 clonal expansions (Fig. 3.6a). This high frequency of clonal expansion is not uncommon after influenza vaccination (Wrammert et al., 2008). The clonal families ranged in size from 2 – 13 detected members and were present in all three populations of interest, with three unique vaccine positive clones that span the IgG and IgA compartments (clone 4, 13 and 21). Clonal expansions containing cells of different isotypes have been reported previously (Horns et al., 2016) and the tendency for BCR sequences to cluster within them by isotype suggests early CSR divergence before continued affinity maturation (Fig. 3.6b). This observation furthers our

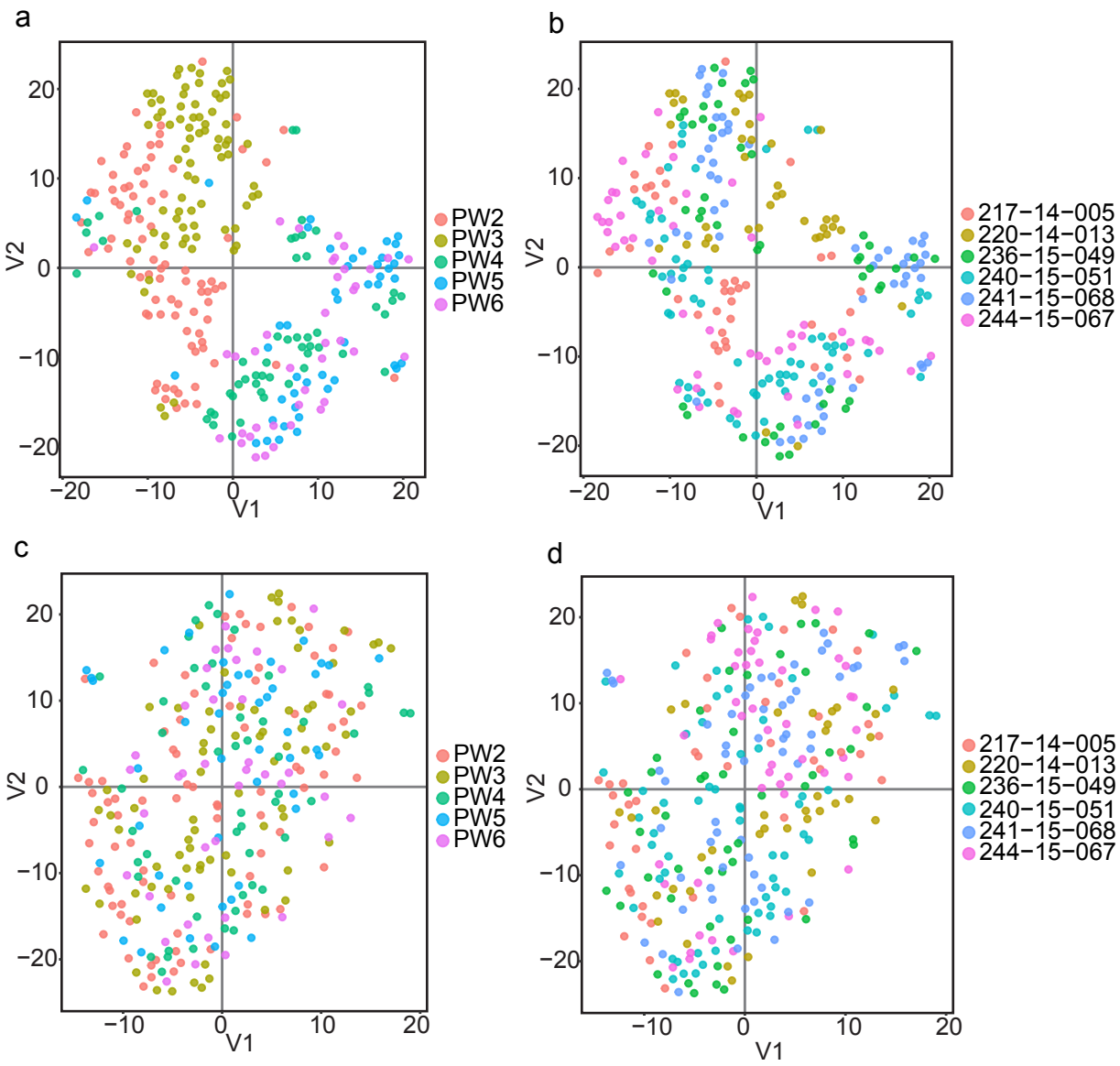


Figure 3.4 Limma batch correction. (**a**) tSNE of data before batch correction labeled by experiment and shows a strong experimental batch effect likely due to the enzyme and oligo modifications between PW2&3 and PW4-6 (n = 295). (**b**) This same tSNE projection colored by patient ID reveals additional undesired clustering. (**c**) the tSNE plot colored by experiment (**d**) and patient ID after limma is used to remove the variai-blity associated with these variables.

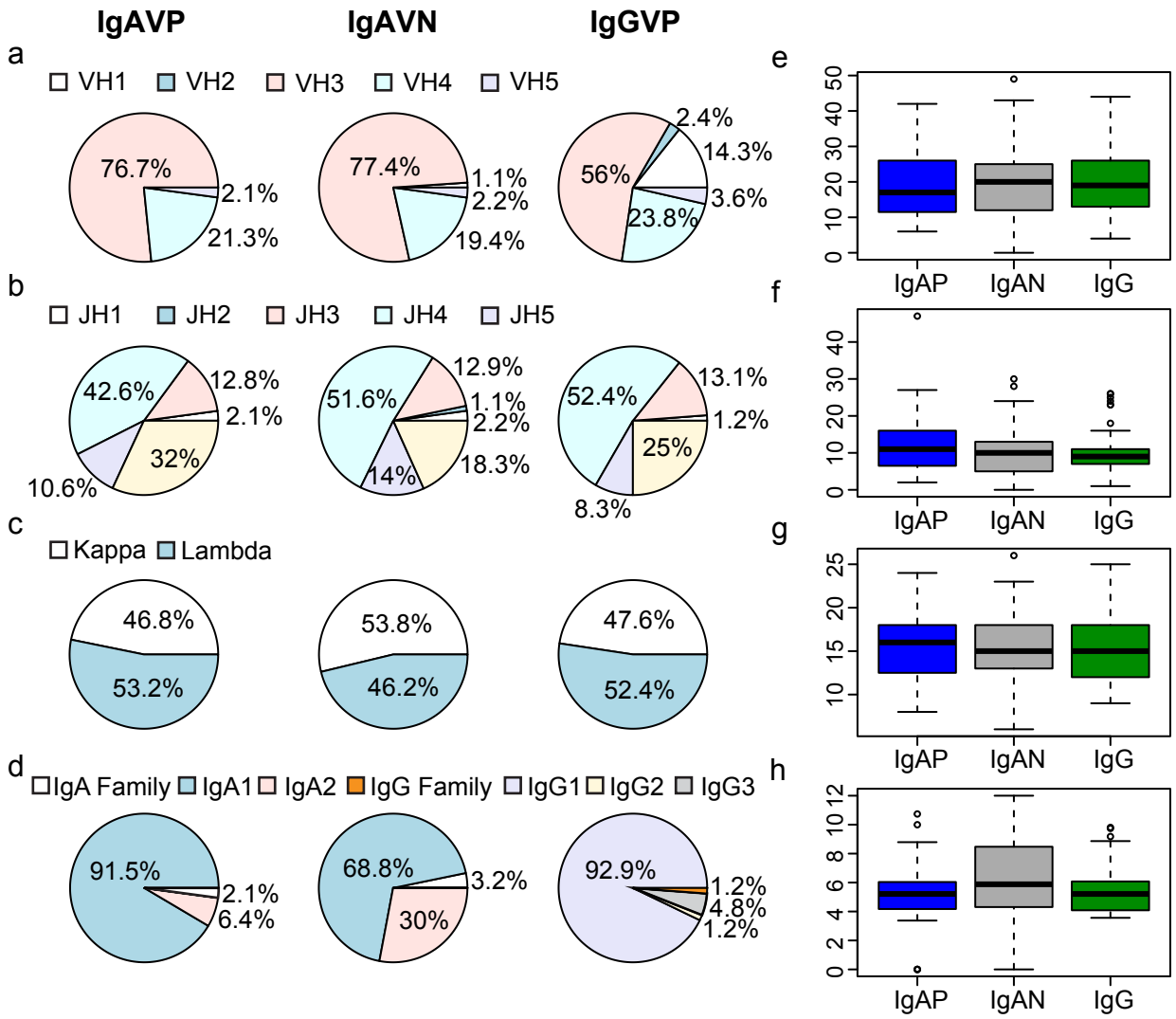


Figure 3.5 Repertoire analysis without clonal expansions. (**a**) Shows V gene family usage, (**b**) J gene usage, (**c**) light chain constant domain usage and (**d**) heavy chain isotype across the three populations. (**e**) Shows the number of heavy chain mutations, (**f**) light chain mutations (p-value = 0.018, one-way ANOVA), (**g**) the heavy chain CDR3 length and (**h**) CDR3 isoelectric point (p-value = 0.051, one-way ANOVA). Median value is displayed in box plots.

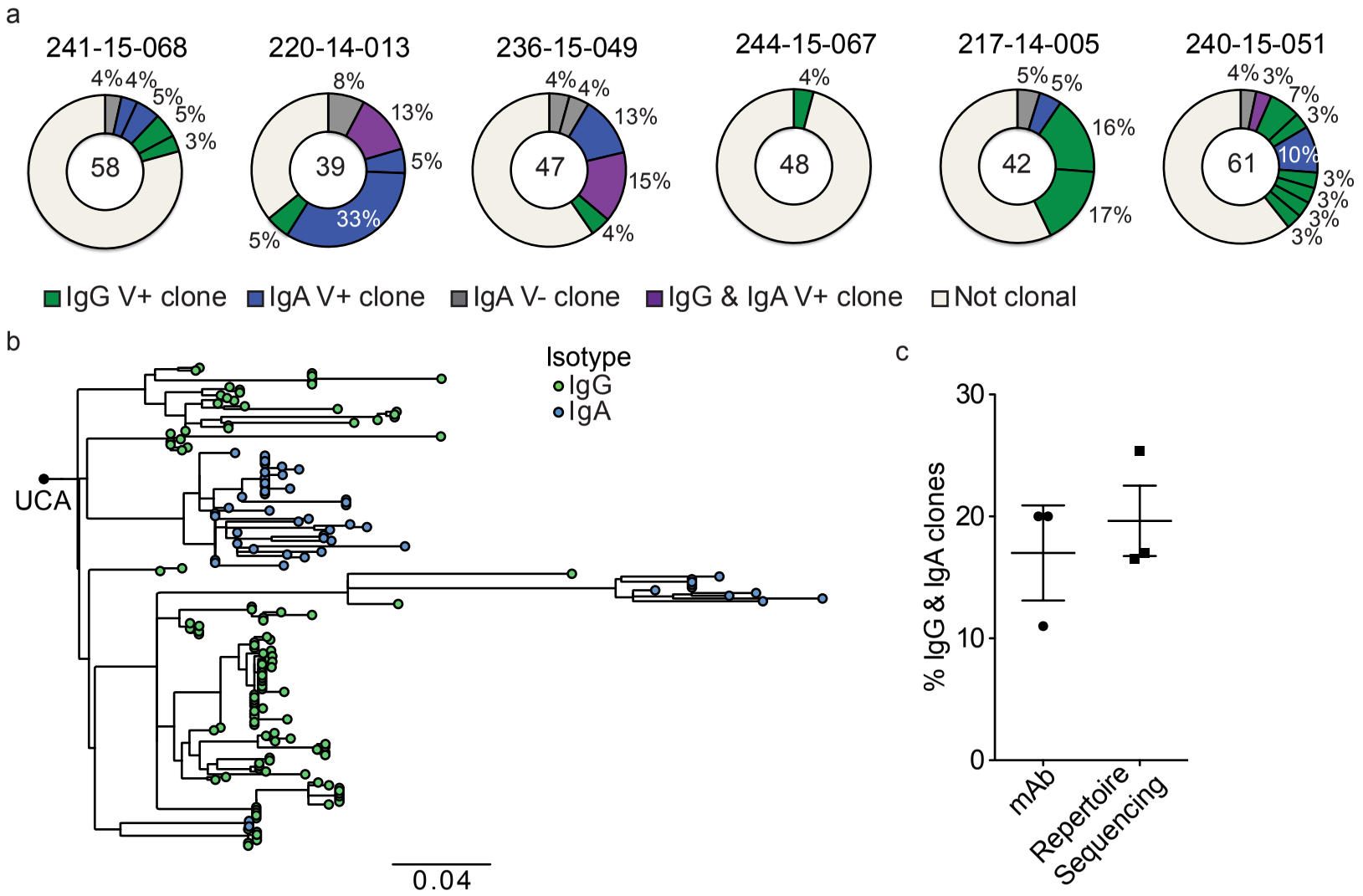


Figure 3.6 Clonal plasmablasts display increased transriptional similarity. (**a**) Clonal expansions are indicated by patient. (**b**) Representative clonal tree from analysis of high throughput repertoire sequencing data. (**c**) Frequency of clones containing both IgG and IgA members, for the three pateints where these clonal expansions were identified. Data mean indicated with line.

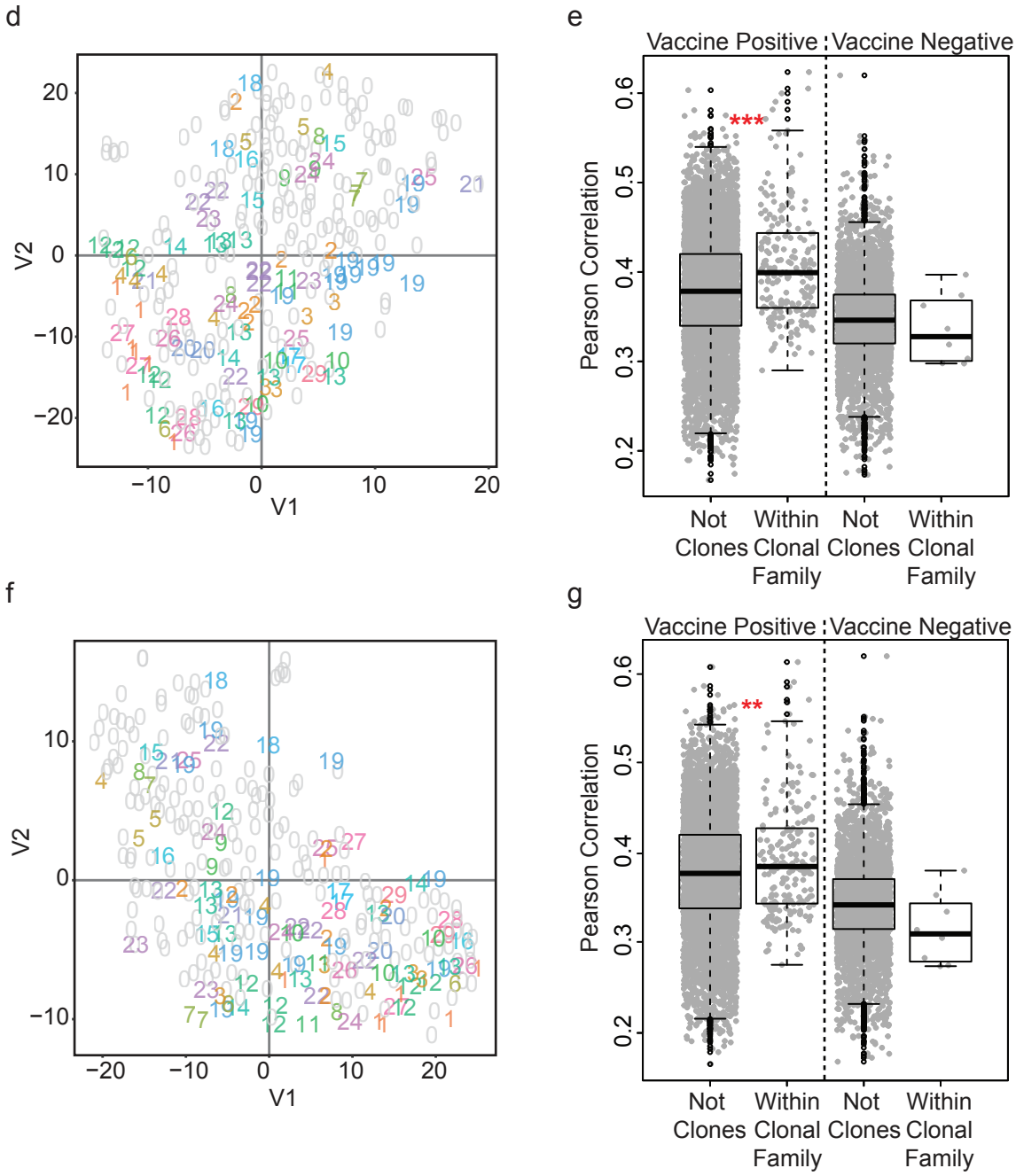


Figure 3.6 Continued. (**d**) tSNE projection of all three populations (n = 295) with clonal families designated with numbers 1-29, and unrelated B cells indicated with grey zeros. (**e**) Pearson correlation coefficients were calculated for all pairwise comparisons of unrelated B cells and between clonal B cells within the same clonal family, for both the vaccine positive (clone n = 87, not clonal n = 108) and vaccine negative (clonal n = 13, not clonal n = 87) compartments. Box plots display median correlation (*** p-value < 2.2e-12, one-sided t-test; Online methods). (**f**) tSNE projection of all three populations after exclusion of Ig genes, with clonal B cells indicated as in (**d**). (**g**) This analysis was repeated after exclusion of all Ig genes (** p-value <1.7e-4, one-sided t-test; Online methods).

interest in comparing IgGVP and IgAVP plasmablasts, as separation supports the possibility of distinct functional identities or activation sites. The 10-20% frequency of mAb clones that contain both IgG and IgA is similar to what we detected with independent high throughput repertoire sequencing studies (16.5%-25.4%) (Fig. 3.6b & c), and supports the shared origin of the IgAVP and IgGVP populations.

The identification of clonal expansions within the IgAVN population was surprising and we have two potential explanations for this. The first would be that the soluble activation signals occurring during antigen specific plasmablast activation were sufficient for BCR independent B cell activation and subsequent proliferation of these cells. Alternatively, this could suggest previous antigen selection events within the IgAVN population resulting in an overall less diverse repertoire. Both of these hypotheses require further exploration in larger clonal datasets.

This unique dataset allowed us to ask if B cells from the same clonal family have more similar transcriptional profiles than unrelated B cells. When clonally related B cells were labeled in the tSNE projection, a preference for related B cells to cluster together was revealed (Fig. 3.6d). To quantitatively evaluate this observation, the Pearson correlation coefficient was calculated for each pair of cells in a clonal family as a way to assess overall transcriptome similarity. The average correlation between clonal family members within the vaccine positive population is significantly higher than those of the unrelated vaccine positive B cells (Fig. 3.6e), indicating increased transcriptional similarity within clonal families. This trend was not detected in the vaccine negative IgA population (Fig. 3.6e). Since B cells clones share the same BCR transcripts, we repeated our analysis in the absence of Ig genes. For clonally related samples in both

the vaccine positive and vaccine negative populations correlations are reduced upon removal of Ig genes, whereas unrelated plasmablast correlations show improvement (Fig. 3.7a & b). The latter improvement is not surprising, as diverse BCR transcripts should in principle increase the heterogeneity of unrelated B cells. Nevertheless, without Ig genes the average correlation within vaccine positive clones is still significantly greater than that of unrelated vaccine positive B cells (Fig. 3.6g) and a preference for related B cells to cluster together remains (Fig. 3.6f). Taken together, although BCR transcripts contribute to transcriptional similarity between clonal plasmablasts, the enhanced transcriptome similarity extends beyond Ig transcripts. As a potential explanation for this we hypothesize that plasmablasts of a clone will have similar immune histories, origins and activation experiences (such as T cell help) that may permanently imprint on the transcriptome and be passed on to clonal progeny. It is intriguing to consider that these potential similarities might drive even single pairs of related B cells to have unique transcriptional profiles.

We then hypothesized that beyond just secreting a particular class of antibody, IgG expressing plasmablasts might be specialized for tissue immunity and those secreting IgA for mucosal immunity. As a means to screen for functional specialization of influenza-induced plasmablasts based on distinct BCR isotype, the gene expression profiles of IgAVP and IgGVP populations were compared. Visualization of these data indicated separate clustering of the two populations (Fig. 3.8a). The most significant differentially expressed genes (DEGs) identified were the genes encoding the BCR constant domains themselves (Fig. 3.8b). In fact, when each single cell was colored according to the expression level of *IGHA1* or *IGHG1* the clusters aligned perfectly with

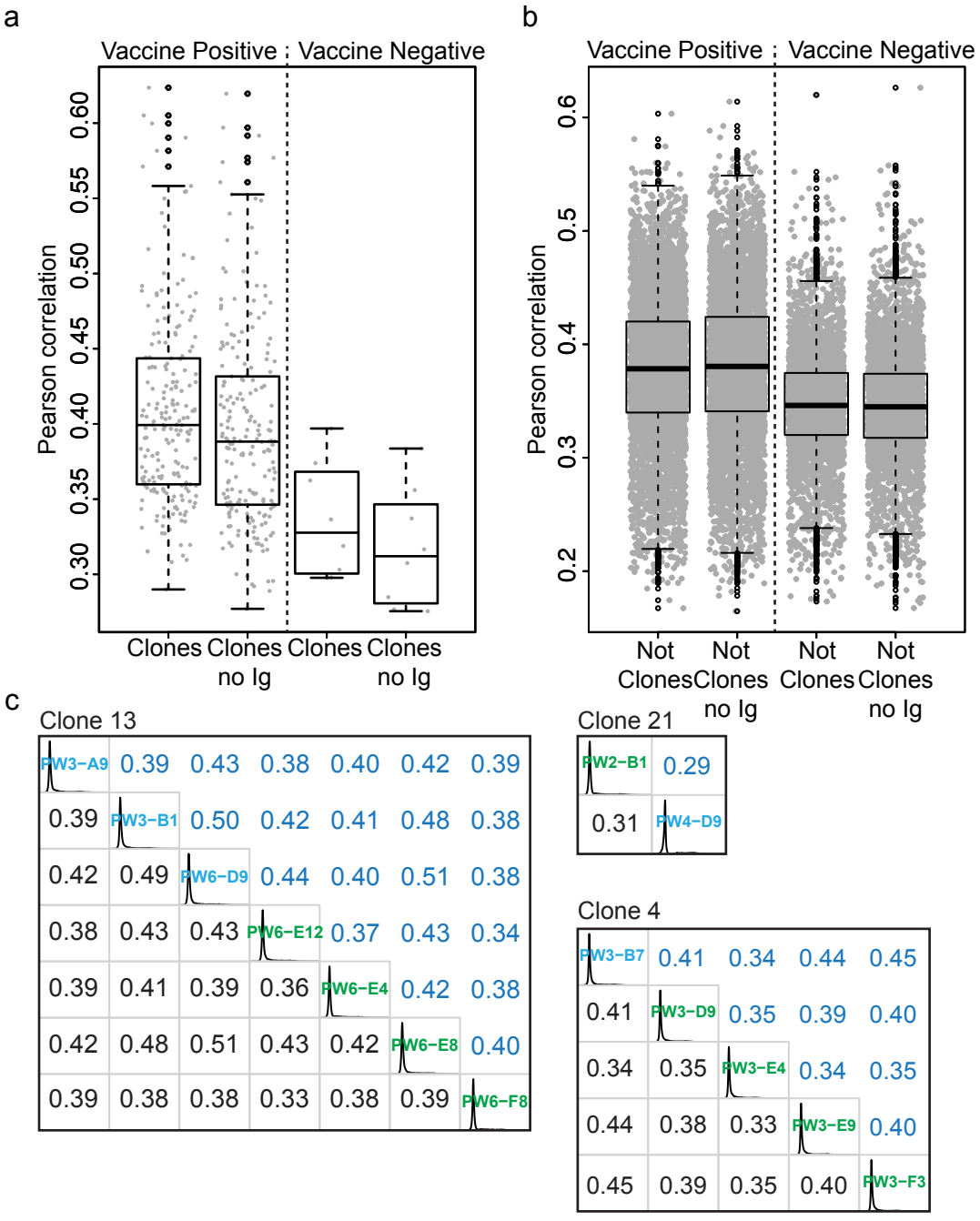


Figure 3.7 The immunoglobulin genes impact the transcriptional similarity of clonally related plasmablasts. (**a**) Pairwise correlations of within vaccine positive (clonal n = 87), or vaccine negative (clonal n = 13), clonal families before and after exclusion of Ig genes. (**b**) Pairwise correlations of unrelated vaccine positive (n = 108) or vaccine negative (n= 87) plasmablasts before and after removal of Ig genes. (**c**) Correlation matrices for the three clones that span the IgGVP and IgAVP compartment. The name of IgGVP cells are in green, and IgAVP are blue. The blue correlation values include all genes, whereas the black correlations are after removal of the Ig constant domain genes. For (**a**) and (**b**) Median correlation is displayed in box plots.

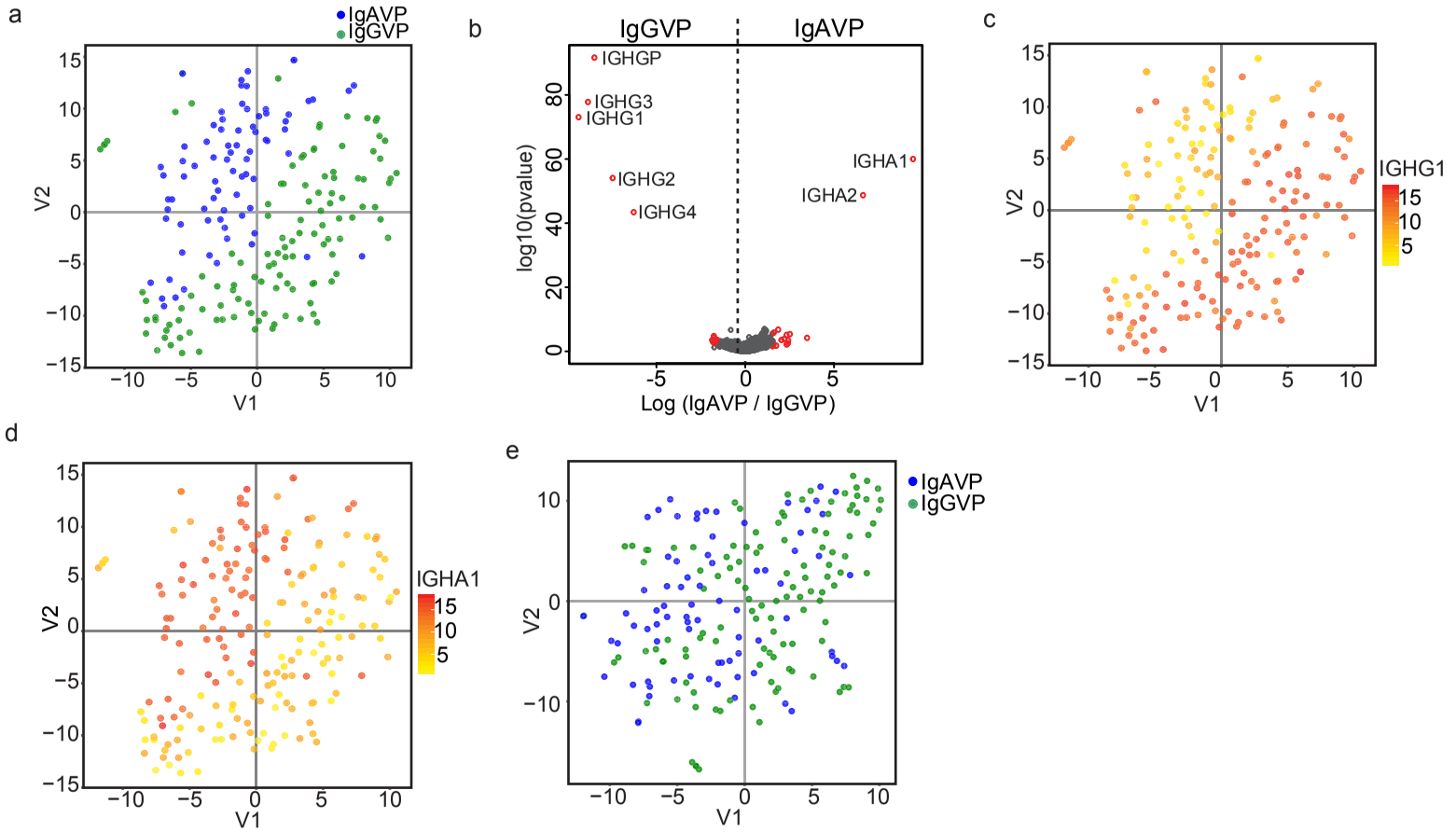


Figure 3.8 BCR constant domain transcripts are sufficient to cause clustering by isotype. (**a**) tSNE projection of vaccine positive cells only (n = 195). (**b**) Volcano plot between IgGVP and IgAVP cells, genes with fold change ≥ to 1.5 with a Benjamini-Hocheberg adjusted p-value ≤ to 0.05 are colored red. (**c**) tSNE projection colored by *IGHG1* gene expression levels, (**d**) and by *IGHA1* gene expression. (**e**) tSNE projection of the vaccine positive cells after Ig-constant genes are excluded.

the assigned isotype annotation (Fig. 3.8c & d). When this analysis was repeated after exclusion of the Ig constant domain genes (Fig. 3.8e), or all Ig genes (Fig. 3.9a), the observed clustering is appreciably diminished, identifying the BCR genes as sufficient to drive transcriptional separation. Notably, non-Ig DEGs alone are capable of directing some clustering (Fig. 3.9a) and grouping (Fig. 3.9b) of cells by isotype. However, our sample size is underpowered to accurately characterize the remaining, subtle, differences for IgA versus IgG class switched plasmablasts and this will need to be explored further with future datasets. As a final measure, we determined that clonal expansions consisting of both IgG and IgA clones were also transcriptionally similar, with and without Ig gene inclusion (Fig. 3.7 c-e), indicating that plasmablasts arising from the same progenitor do not differentiate further after class switching to IgG or IgA. From these analyses we conclude that, besides the BCR genes themselves, peripheral blood IgG or IgA plasmablasts induced by influenza vaccination have highly similar transcriptional profiles.

In order to determine if differences exist between antigen induced and steady state plasmablasts two analyses were performed. The first was to compare the IgAVP and IgAVN samples only. The tSNE projection of the IgA cells suggests transcriptional segregation between these two populations (Fig. 3.10a), which could indicate different functional attributes. In total, 49 DEGs with an adjusted p-value ≤ 0.05 and a fold-change ≥ 1.5 were identified between the IgAVP and IgAVN populations (Fig. 3.10b). The significantly increased expressions of *IGHA2, VH3-23, VH3-7* and *IGKC* within the IgAVN population, and enrichment of *IGLJ2* in the IgAVP population, are confirmed by repertoire gene usage frequencies (Fig. 3.10b & c). The increased frequency of *IGHA2*

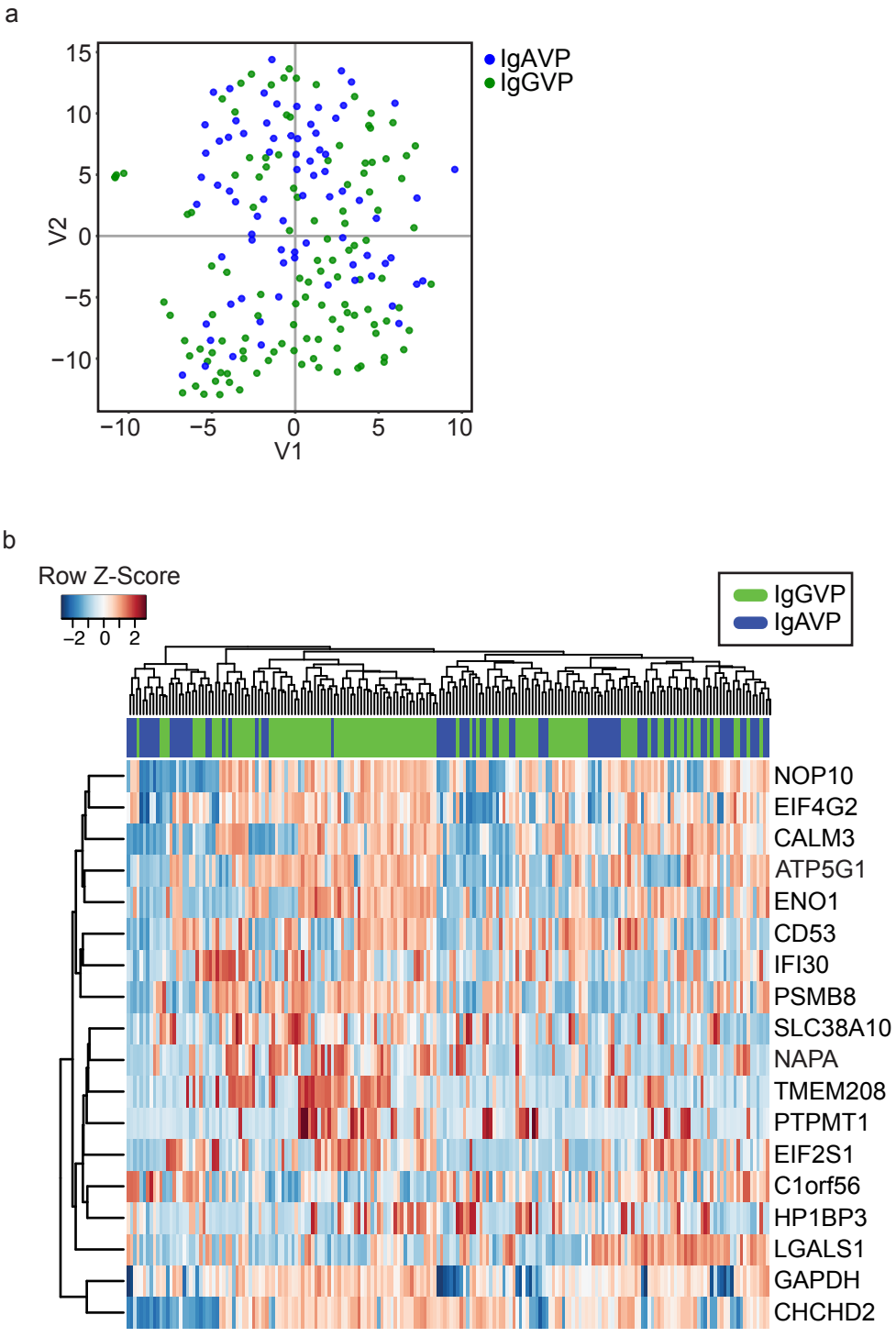


Figure 3.9 Other genes contribute to plasmablast clustering by isotype. (**a**) tSNE projection of vaccine positive cells (n = 195) after removal of all Ig genes. (**b**) heat-map of non-Ig DEGs with a log fold ≥ 1.5 and a Benjamini-Hoeschberg adjusted p-value (two-sided t-test) ≤ 0.05 on IgGVP and IgAVP cells.

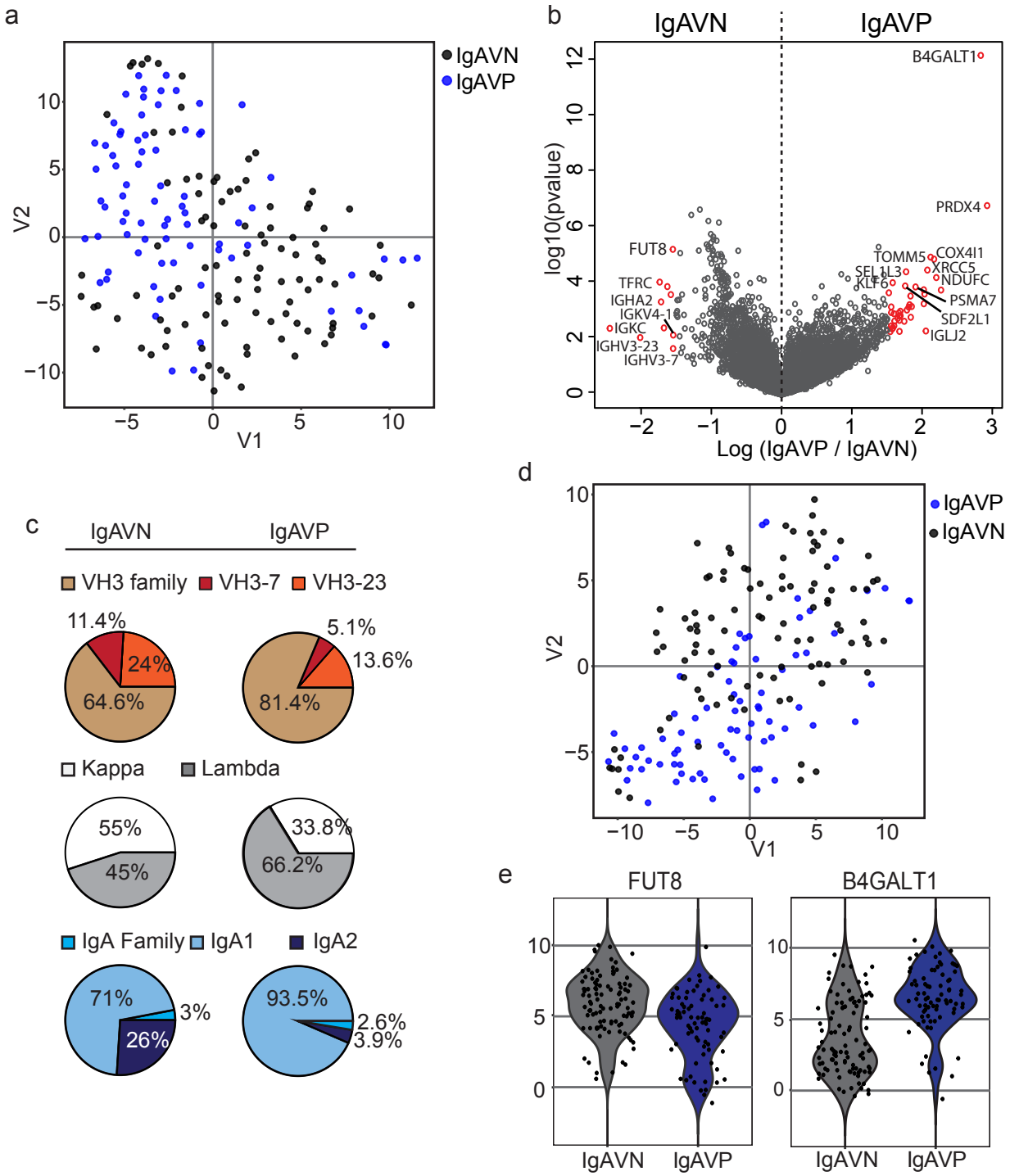


Figure 3.10 IgA plasmablasts cluster by their ability to bind the vaccine. (**a**) tSNE projection of the IgA plasmablasts (n = 177). (**b**) Volcano plot showing the differential expression of all genes between IgA vaccine negative (IgAVN) and IgA vaccine positive (IgAVP) plasmablasts. Genes with log fold ≥ to 1.5 with a Benjamini-Hocheberg adjusted p-value (two-sided t-test) ≤ to 0.05 are colored red. Some of the most significant genes are labeled. (**c**) Pie charts showing repertoire usage frequency between IgAVN and IgAVP plasmablasts, as validation of some identified DEGs. (**d**) IgA tSNE repeated after removal of all Ig-genes. (**e**) Violin plots of *FUT8* and *B4GALT1* between IgAVN and IgAVP plasmablasts.

usage within IgAVN cells supports their contribution to antibody responses at mucosal sites, as this IgA subtype is more resistant to protease degradation (Spencer and Sollid, 2016). In order to assess the impact of these genes on overall population clustering the tSNE analysis was repeated without any Ig genes, and similar clustering was identified (Fig. 3.10d). This indicates transcriptional differences exist beyond repertoire usage. Interestingly, the two most significant DEGs between IgAVP and IgAVN have been previously implicated in BCR glycosylation; *FUT8* and *B4GALT1* (Fig. 3.10b & e).

Our second analysis compared the total vaccine positive samples with the IgAVN population, after exclusion of all Ig genes. The tSNE projection confirms separation based on specificity, showing a preference for the IgGVP and IgAVP cells to cluster together, and away from the IgAVN cells (Fig. 3.11a). A larger number of DEGs were identified in this comparison, perhaps due to the increased sample size, including *FUT8* and *B4GALT1* (Fig. 3.11c & d). These genes are known to shape a humoral immune response through alterations to Fc-receptor binding and complement activation (Basset et al., 1999; Gomes et al., 2008; Hodoniczky et al., 2005; Keusch et al., 1998; Selman et al., 2012; Shade and Anthony, 2013; Shields et al., 2002; Woof and Mestecky, 2005; Yamane-Ohnuki et al., 2004). The differential expression of these enzymes between vaccine positive and vaccine negative cells, suggests plasmablast BCR specificity directs transcriptional differences that can ascribe key effector properties to the antibodies secreted by that cell.

FOXP1 was also identified as differentially expressed between total vaccine binding and vaccine negative plasmablasts (Fig. 3.11c & 3.12a). This transcription factor is known to repress terminal differentiation into antibody secreting cells through

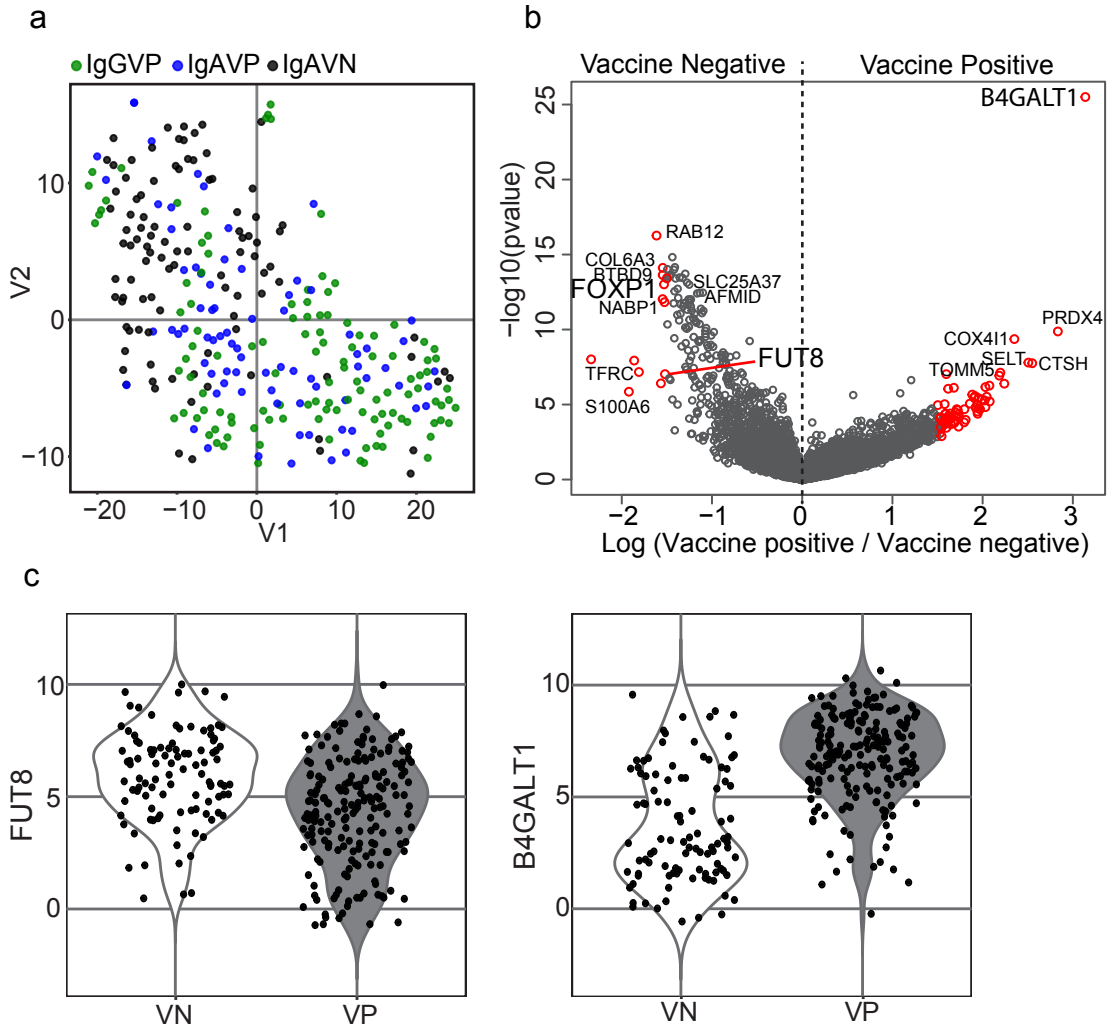


Figure 3.11 Total plasmablasts cluster by their ability to bind the vaccine. (**a**) tSNE projection of all three populations after removal if Ig genes (n = 295). (**b**) Volcano plot showing DEGs between total vaccine positive and vaccine negative cells. (**c**) Violin plots showing expression of B4GALT1 and FUT8 in the total vaccine positive (VP) versus vaccine negative (VN) plasmablasts.

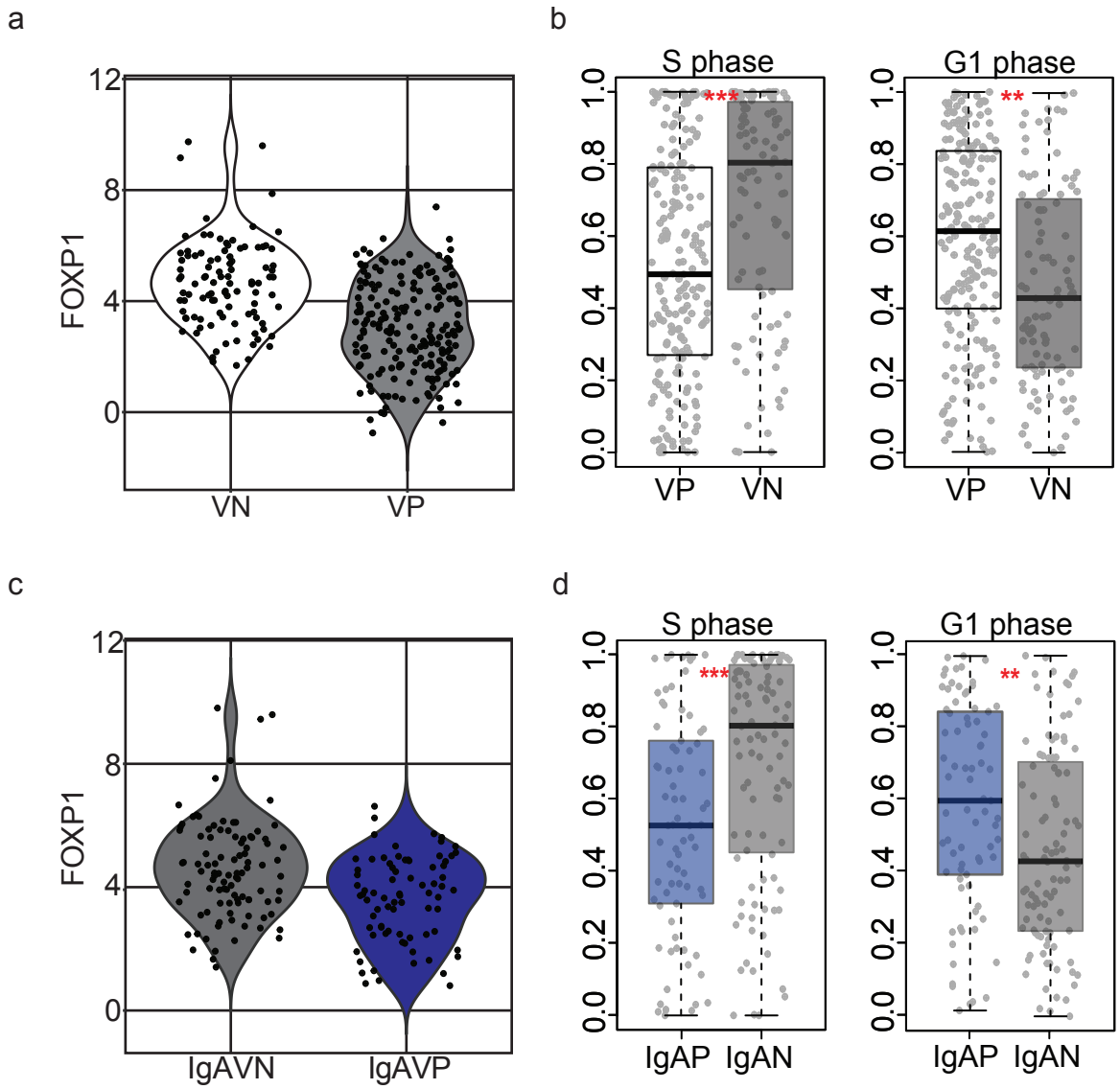


Figure 3.12 *FOXP1* expression in IgAVN cells blocks cell cycle arrest in G1 phase. (**a**) Violin plot showing *FOXP1* expression between total vaccine positive (VP) and vaccine negative (VN) plasmablasts. (**b**) Box plots showing differences in frequency of each cell cycle stage between IgAVP and IgAVN plasmablasts. Box plots display median cell cycle score. (p-values < 1.4e-4 ** and < 9.4e-6 ***, two-sided t-test). (**c**) Violin plot for *FOXP1* expression between IgA vaccine negative (IgAVN) and IgA vaccine positive (IgAVP) plasmablasts. (**d**) Same analysis as in b except applied to IgAVN and IgAVP plasmablasts (p-values <4e-3** & <3e-4***, two-sided t-test).

repression of *PRDM1*, *IFR4* and *XBP1*, the master transcription factors of the plasmablast fate (Keimpema et al., 2015). The increased expression of *FOXP1* within the vaccine negative plasmablast population may suggest up-regulation of pathways designed to prevent terminal differentiation into traditional antigen-induced plasmablasts. To explore this further we performed cell cycle analysis, as G1 phase arrest is an accepted feature of terminally differentiated antibody secreting cells (Chen-Kiang, 2003; Ezhevsky et al., 1996). Transcriptome cell-cycle analysis revealed an increased frequency of vaccine positive cells assigned to the G1 phase, versus the IgAVN compartment displaying increased assignment to the S phase (Fig. 3.12b). This trend was also true between IgAVP and IgAVN cells (Fig. 3.12d). IgAVP cells express lower levels of *FOXP1* than the IgAVN cells (Fig. 3.12c) although the measured fold change fell below our requirements to be identified as a DEG between these populations. This suggests that the vaccine negative plasmablasts have prevented terminal differentiation through up-regulation of *FOXP1*, to allow for continued cellular division, perhaps as a mechanism for population renewal and maintenance. Interestingly, differences in BCR signaling or B cell activation pathways between IgAVP and IgAVN plasmablasts could not be detected (Fig. 3.13 & 3.14). Altogether, these data supports the possibility of an antigen-induced gene expression profile that, if explored further, could be used to identify vaccine specific cells directly from their scRNA-seq data without the need for time-exhaustive receptor generation and testing.

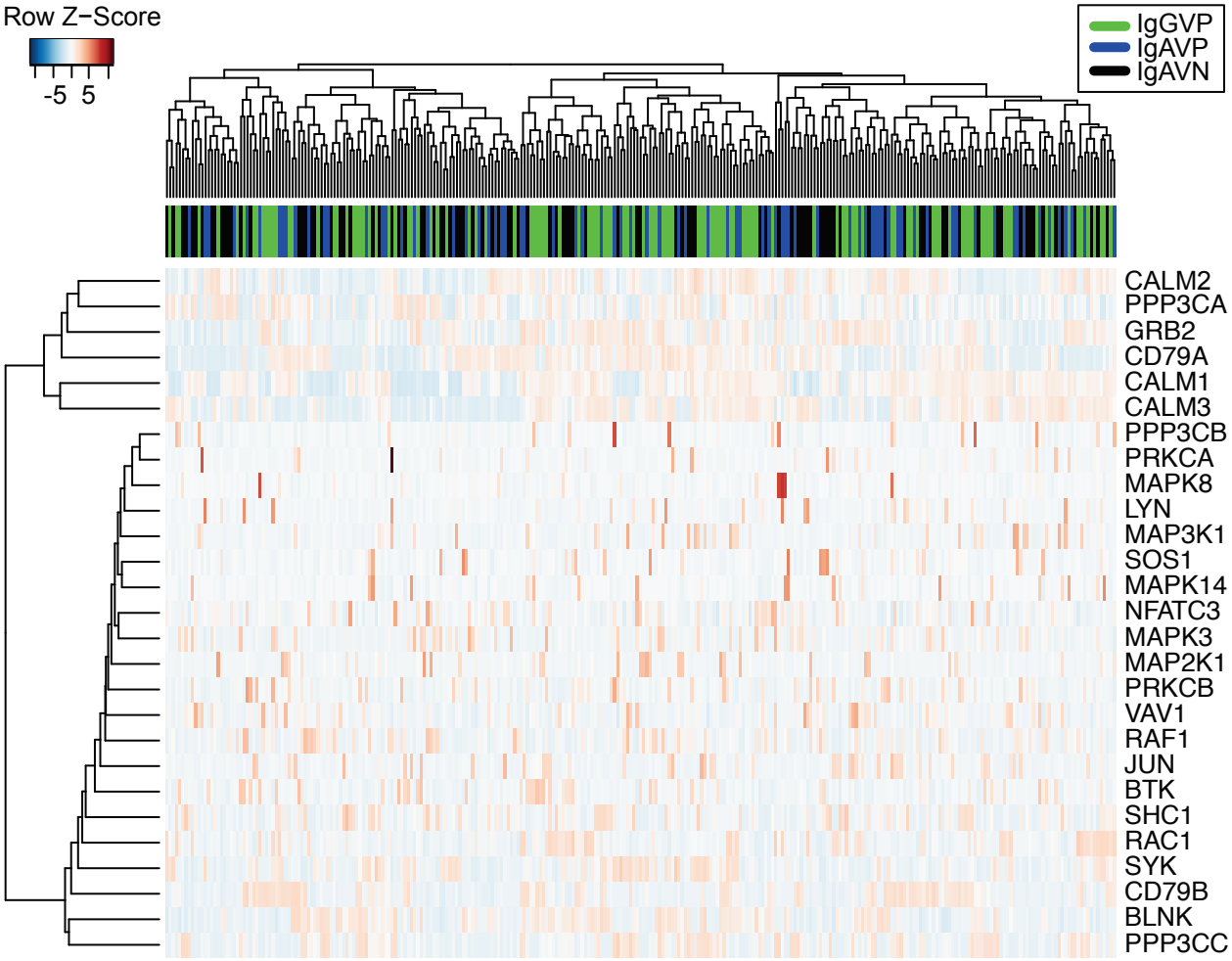


Figure 3.13 No differences in BCR signaling were identified between the three populations of interest. Heatmap shows gene expression scaled by row for a list of genes assocaited with BCR signaling.

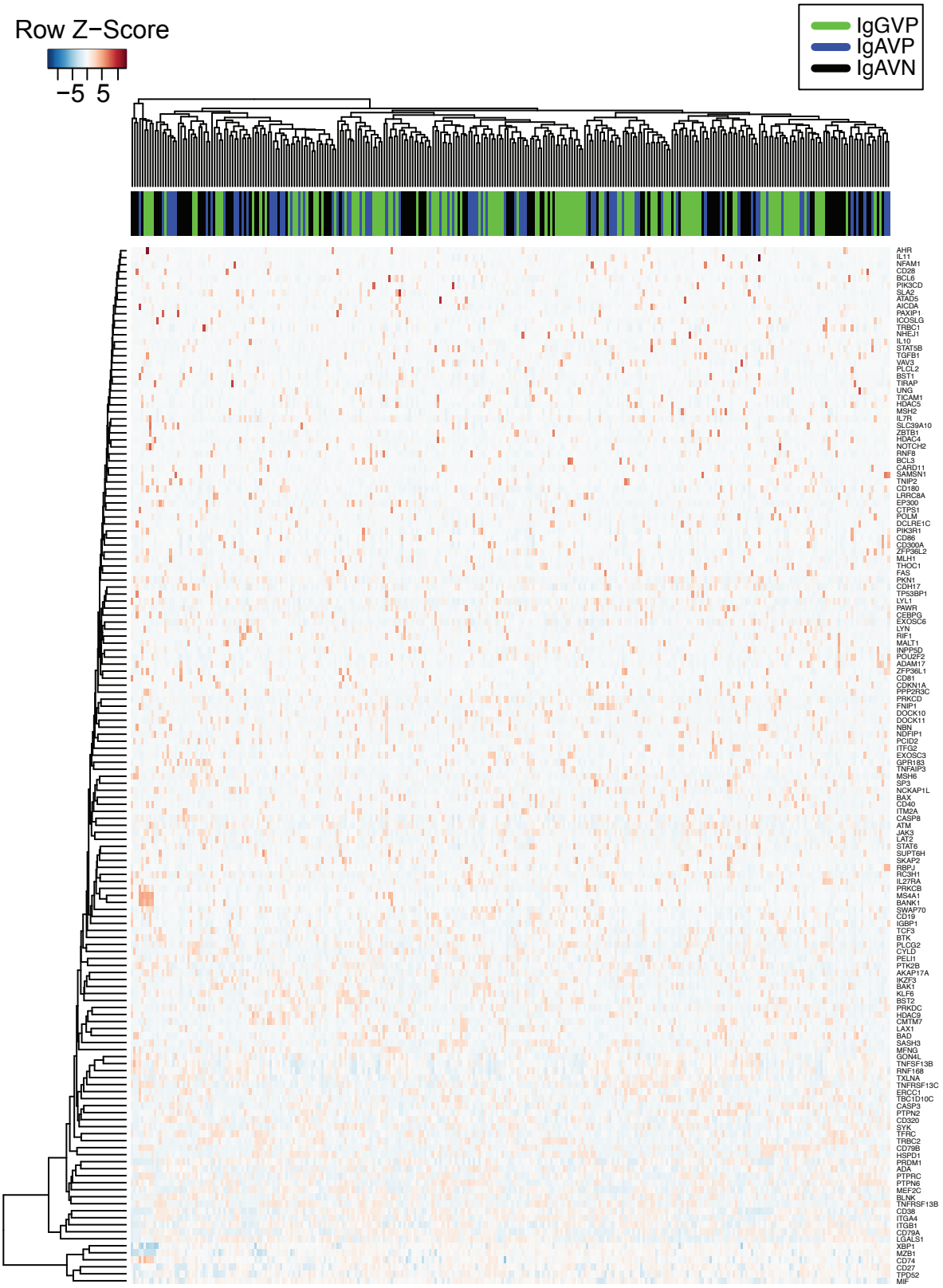


Figure 3.14 No differences in B cell activation were identified between the three populations of interest. Heatmap shows gene expression scaled by row for a list of genes assocaited with B cell activation.

3.3 Discussion

In this study, we applied Spec-seq to peripheral blood plasmablasts and found that the BCR transcripts are the major drivers of population heterogeneity. However, transcriptional differences beyond the BCR, including genes associated with Fc-glycosylation regulation and cell cycle progression were identified between vaccine positive and vaccine negative plasmablasts (Fig. 3.15).

Specifically, our data shows that differential expression of *B4GALT1*, *FUT8* and *FOXP1* is directly related to BCR recognition of the vaccine. With an increased frequency of vaccine positive cells arresting in the G1 phase of the cell cycle and expressing high levels of *B4GALT1* which may facilitate the production of antibodies with enhanced ADCC potential. Whereas, the vaccine negative plasmablasts displayed increased expression of *FOXP1* and *FUT8* suggesting blockade of terminal ASC differentiation and the production of anti-inflammatory antibodies. As expected with increased *FOXP1* expression a larger frequency of vaccine negative plasmablasts were assigned to the S phase of the cell cycle, indicating DNA replication and continued cellular division (Chen-Kiang, 2003; Ezhevsky et al., 1996). As the IgAVN cells are hypothesized to be sentinel cells recirculating from mucosal sites it is plausible that proliferation is necessary to maintain peripheral blood population numbers. Both of these distinctions need further exploration, but suggest the IgAVN plasmablasts may perform a homeostatic function through production of antibodies lacking the inflammatory properties associated with antigen-induced antibodies and through proliferative renewal within the peripheral blood compartment. Further characterization

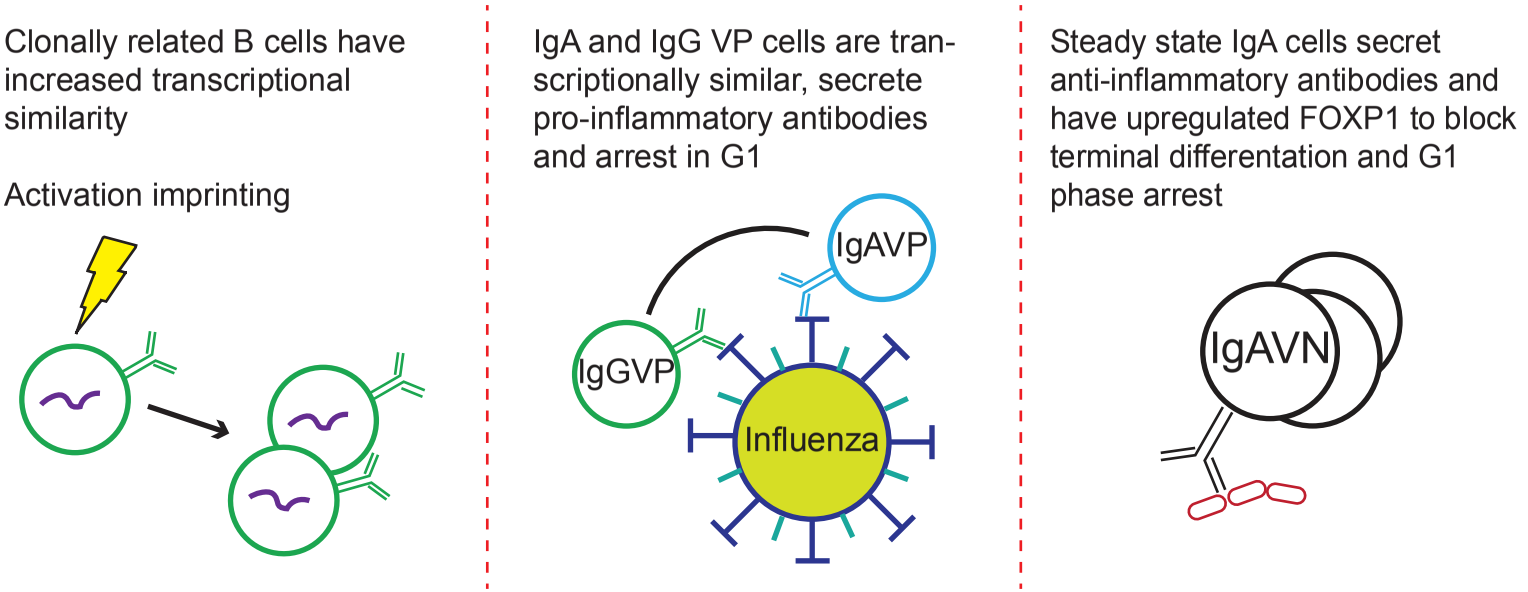


Figure 3.15 Conclusions. Our results suggest three main conclusions. Firstly, they show that clonally related B cells have enhanced transcriptional similarity beyond their shared BCR genes suggesting possible original B cell activation imprinting. Secondly, our comparison of IgA vaccine positive (IgAVP) and IgG vaccine positive (IgGVP) plasmablasts identified a high level of transcriptional similarity outside of the BCR constant domain transcripts. Thirdly, the comparison of the vaccine positive cells against the IgA vaccine negative (IgAVN) cells, identified differences in enzymes responsible for antibody glycosylations and cell cycle control. Overall, this suggests that the IgAVN population is transcriptionally distinct from the antigen-specific vaccine induced population.

of these antibodies to determine antigen specificity will be very interesting and could finally give concrete evidence for their role in regulating the microbiome.

These findings may also suggest that once a plasmablast reaches the peripheral blood it is specialized for peripheral blood antibody production, and that future characterization of mucosally localized ASCs is necessary to better understand how to induce mucosal protection. Assessment of the incorrect niche was also partially responsible for the contradictory results of previous studies on mucosal vaccine efficacy (Barría et al., 2013; Nakaya et al., 2011; Ohmit et al., 2011; Weinberg et al., 2010). These studies imply that although peripheral blood ASCs have a shared origin with mucosal ASCs (Iversen et al., 2017), transcriptional profiles likely reflect the currently occupied niche. Further characterization of mucosally localized ASCs is therefore instrumental in improving the immunogenicity of mucosal vaccines.

Full-length antibody sequences were also obtained through Spec-seq, which enabled us to perform the first-ever analysis of the singe cell transcriptome of clonally related B cells. Unexpectedly, our analysis suggests that clones share increased transcriptional similarity beyond the expected Ig-genes. These data raise the tantalizing possibility that original activation events are maintained within the transcriptome, and should be explored further with datasets containing larger clonal expansions.

3.4 Contributions

K.E.N carried out the majority of experiments and data analysis. M.H performed antibody cloning. J.L assisted with data analysis. K.T.R coordinated patient recruitment, sample collection and assisted in experiments. C.H provided key guidance,

experimental assistance and manuscript editing. M.C.V, K.K and S.C performed analysis of high throughput repertoire sequencing data. A.A.K directed the computational analyses. P.C.W conceived of the project and directed all experiments. K.E.N, A.A.K and P.C.W wrote the manuscript.

4. Discussion

4.1 Overall implications of this work

Spec-seq opens new doors for the characterization of transcriptional heterogeneity in the context of antigen specificity. This is particularly important for cells of the adaptive immune system where unique somatically rearranged receptors identify cells of distinct antigen specificities, which equates to distinct effector functions. Many aspects of B cell biology are linked to BCR specificity and this approach will be instrumental to furthering this field.

Specifically, I would like to mention two areas I would be very excited to see utilize this approach. Firstly, would be to look at the process of affinity maturation. As mentioned in the introduction this process involves repeating rounds of BCR SHM and testing the ability of the mutated BCR to bind and process antigen for peptide presentation (Fig. 3.1a). Spec-seq could facilitate a link between transcriptional profiling of B cells and the antigen affinity of the BCR. Clonally related B cells can now be identified in conjunction with scRNA-seq, and it is possible that disparate transcriptional profiles could explain the clonal dominance sometimes observed in an ongoing germinal center response (Tas et al., 2016). Relatedly, I would like to see Spec-seq applied to the process of lineage commitment that occurs when B cells leave the germinal center (Fig. 3.1a). This approach may be able to finally explain what transcriptional changes allow clonally related B cells to leave the germinal center and be destined for different compartments; such as tissue resident memory or long lived plasma cells. Spec-seq could also be easily modified for exploration of T cells. This would not only further our

understanding of this diverse branch of the adaptive immune system, but also allow for interrogation of intricate antigen-specific populations such as tumor infiltrating T cells.

Our successful application of Spec-seq to human peripheral blood antibody secreting cells has expanded our understanding of this population and created interesting avenues for future studies. Although novel, the identification of BCR transcripts as the dominant mediator of plasmablast subpopulations is not a surprising result. These cells specialize in extreme production of antibodies, which initiates a strong unfolded protein response, the subsequent accumulation of reactive oxygen species, and eventual cell death (Shaffer et al., 2004). Due to this specialization, mRNA for the BCR protein will be continuously synthesized and highly over-represented. This high expression explains why variable V(D)J-C usage between B cells is a large source of transcriptional heterogeneity, and our analysis supports this with the identification of clusters specifically associated with expression of the selected constant domain (Fig. 3.8). Our analysis found no further significant changes in the gene expression profiles of cells expressing IgG versus IgA BCRs, and therefore suggests that CSR does not induce cellular functional specialization.

These conclusions have implications for the field of vaccine immunology, as our findings suggest that peripheral blood vaccine positive IgA plasmablasts are more transcriptionally similar to vaccine positive IgG plasmablasts than vaccine negative IgA plasmablasts. Therefore, it seems unlikely that simply inducing a larger frequency of IgA plasmablasts in the peripheral blood following vaccination would improve mucosal protection and instead of isotype, the field needs to continue to focus on other ways of inducing a mucosally localized protective population of ASCs (Ichinohe et al., 2007; van

Riet et al., 2012; Sundararajan et al., 2015; Wang et al., 2010). The main avenues for this would be through the incorporation of distinct vaccine adjuvants or altering the route of vaccine administration. Therefore, additional studies are still needed to understand how to best induce mucosally localized protection through vaccines.

Our ability to measure receptor binding to the influenza vaccine itself provided unique insight into gene signatures associated with antigen specific antibody production. These results identified differential expression of *B4GALT1* and *FUT8,* enzymes that are known to regulate antibody glycosylation. This is interesting because different glycosylation moieties on the Fc-domain are known to alter the effector properties of antibodies (Hayes et al., 2014; Maverakis et al., 2015). Specifically, the increased expression of *B4GALT1* identified within vaccine positive plasmablasts suggests the addition of terminal galactose moieties, which enhance the pro-inflammatory effector properties of the secreted antibody. Although these enzymes have been studied in bulk populations, our work is the first to identify them at the single cell level and the first to tie their expression directly to antigen reactive antibodies. Further work will need to be performed to confirm that increased enzyme expression equates to differential glycosylation of the secreted antibody protein.

Plasmablasts have long been considered terminally differentiated antibody secreting cells with limited complexity. Our data disproves that, with evidence of unexpected levels of transcriptional specialization within this population. It also provides evidence of a possible new, distinct, subset of plasmablasts; the steady-state peripheral blood plasmablasts. The transcriptional profiles of these cells suggest that they secrete anti-inflammatory antibodies, remain active in the cell cycle and have up-regulated

pathways known to prevent terminal differentiation. The overall significance of these cells remains to be elucidated. A key next step in characterization will be exploring BCR specificity to determine if these cells recognize microbial antigens, as this could solidify a connection to the mucosally localized, microbiome-regulating IgA ASCs.

We were surprised that our analysis did not reveal expression of any mucosal homing receptors (Fig. 4.1). This could be explained by the disconnect between RNA transcripts and protein levels, making it possible for specific homing receptors expressed at the protein level to be missed in the RNA gene expression profile. It is also conceivable that the expression of alternative niche homing receptors is lost upon localization into the peripheral blood and may only be up-regulated during a re-location event. Therefore, identification of mucosal homing transcriptional signatures will likely require direct Spec-seq application to a mucosally localized cell population.

4.2 Current and future directions

We are currently exploring assays to confirm that the steady state plasmablasts secrete fucosylated antibodies whereas plasmablasts induced after influenza vaccination secrete galactose-modified antibodies. Although, evidence in the field has already shown that influenza specific antibodies in the peripheral blood have increased galactose modifications after vaccination and steady-state antibodies have increased fucosylation (Selman et al., 2012), the direct connection between these serum antibodies and peripheral blood plasmablasts has never been made. The origin of these antibodies could have been from long-lived plasma cells that are known to contribute significantly to serum antibody titers. However, our data suggests these antibodies are

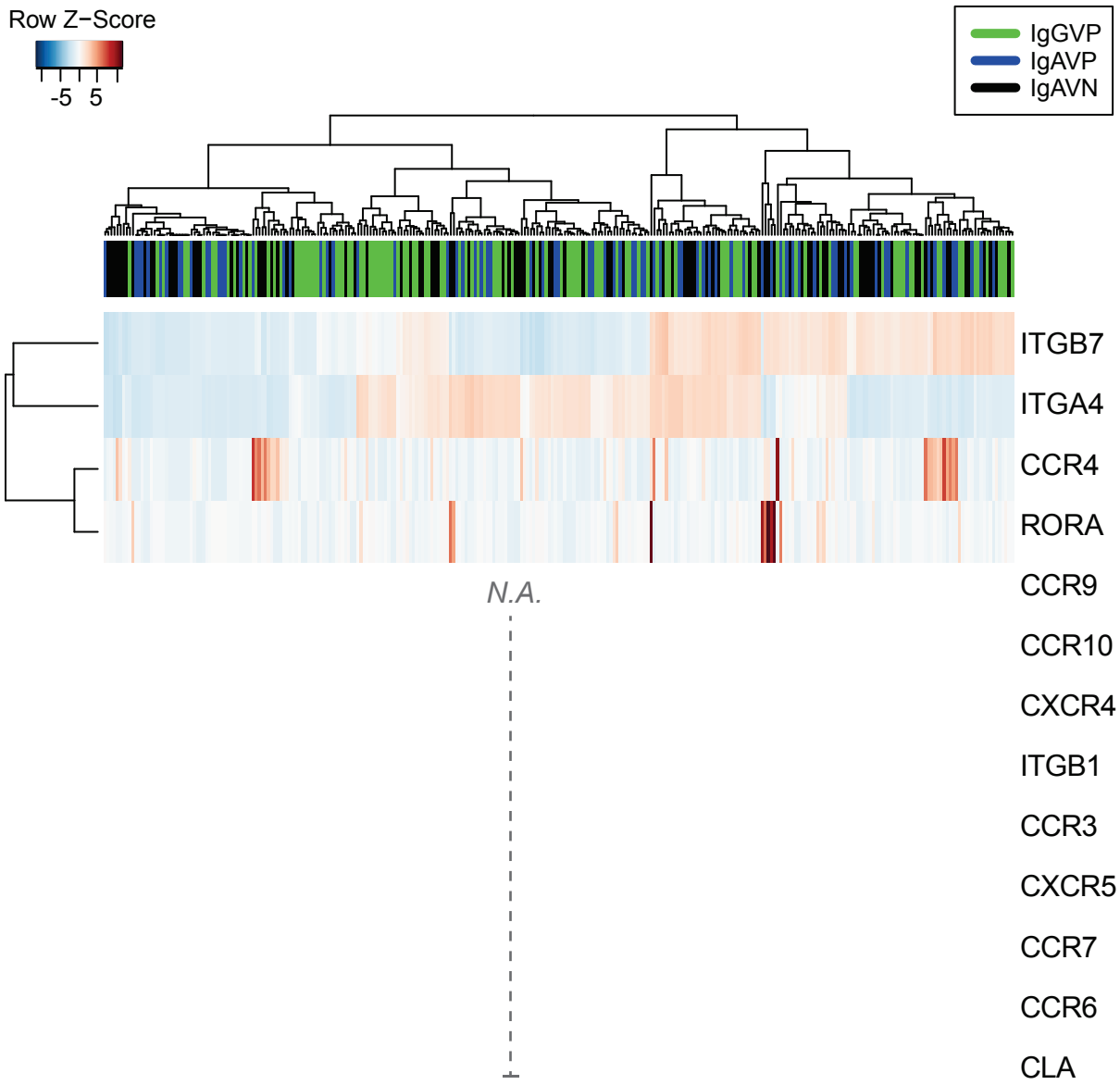


Figure 4.1 No reliable expression of mucosal homing genes was identified in the three populations of interest. Heatmap of genes assocaited with mucosal homing for all three cell populations; IgAVP (blue), IgAVN (black) and IgGVP (green). Genes that were not present in the filtered dataset are represented in the heatmap as *N.A.*

at least in part derived from the plasmablast population. In addition to validating the differential glycosylations of antibodies we will also assess cell cycle activity with flow cytometry to validate the continued self-renewal of the vaccine negative plasmablast compartment.

As a next direction for this project we set out to determine if transcriptional differences could be identified between peripheral blood plasmablasts induced by the seasonal vaccination and natural influenza infection. We hypothesized that infection induced plasmablasts would have evidence of their mucosal origin and any transcriptional specialization associated with that. We also hypothesized the plasmablasts induced by infection may have evidence of ongoing viral infection or perhaps anti-viral machinery to prevent that. To assess this we collected peripheral blood from three patients confirmed to have ongoing H1N1 influenza infections in 2015-16 and three age and gender matched 2016-17 influenza quadrivalent vaccine recipients. The plasmablasts were isolated with the same markers as our earlier vaccination studies. The plates were processed with the same Spec-seq protocol, except for one significant change. During the troubleshooting process for the original Spec-seq study a mistake was made in the catch buffer, where an excess of dNTPs were added. Exploration of this protocol alteration revealed that the dNTP concentration recommended by Smart-seq2 yielded significantly improved data (Fig. 4.2). Therefore, all samples included in this new study were sorted into the originally recommended dNTP concentration.

352 single cells (194 from infection and 158 from vaccination) with a successfully amplified BCR heavy and light chain were sequenced for this study. The libraries were

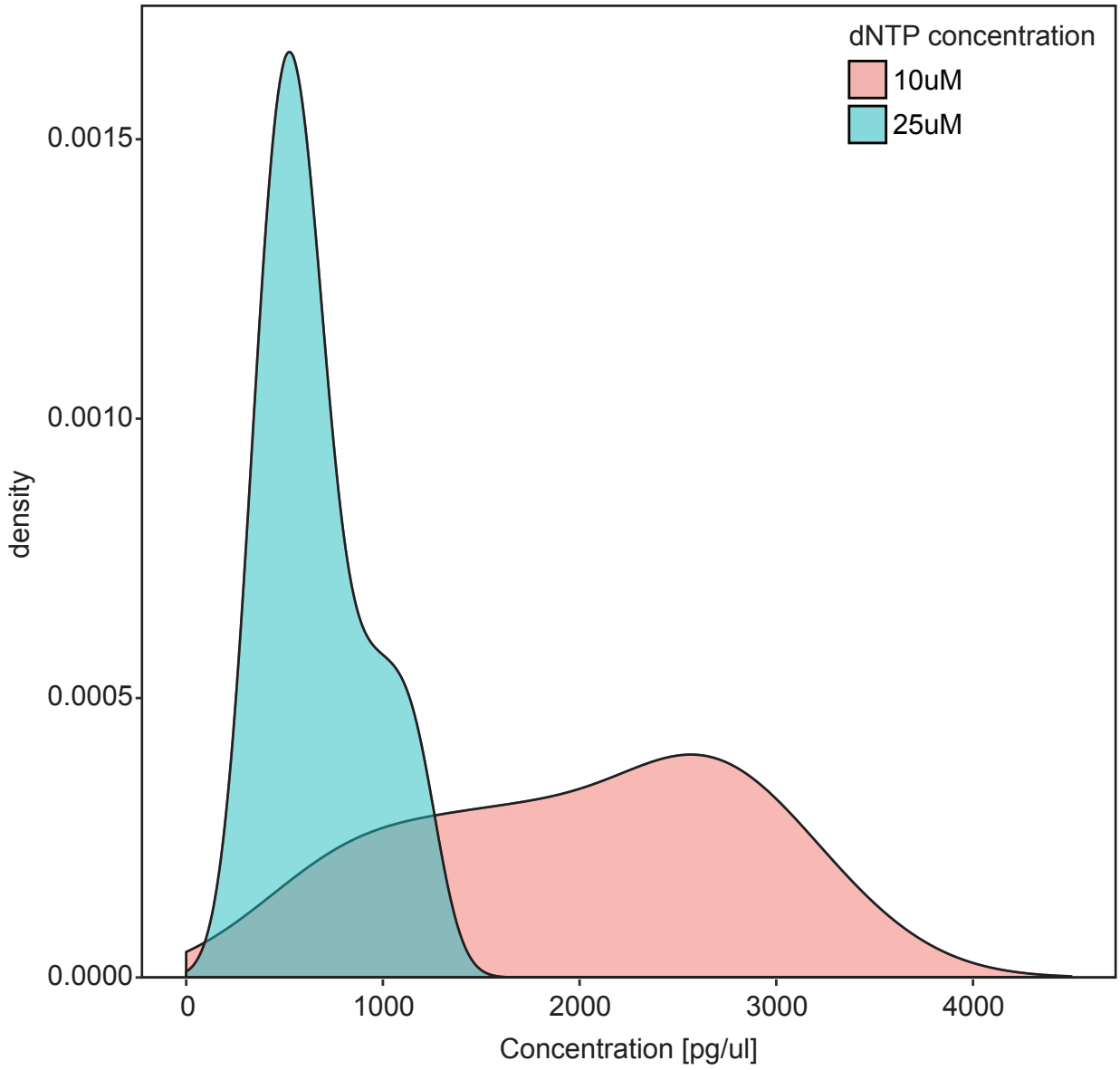


Figure 4.2 dNTP concentration comparison. Density plot shows effect of varried dNTP concentration on cDNA yield. Day 7 plasmablasts were sorted from a vaccinated patient into plates containing different cDNA concentrations to perform this comparision.The recommended 10uM concentration gives significantly improved results over the previously utilized 25uM. cDNA concentration was obtained from Agilent Bioanalyzer high sensitivity DNA chips.

pooled at 3 nM each and each set of 96 was sequenced in a single lane of the HiSeq2500 machine at the University of Chicago Genomics facility. As described in Chapter 2, careful attention was paid to sample organization across plates in order to simplify future assessment of batch effects. These data will be analyzed with our established analysis pipeline, discussed in Chapter 2 and covered in detail in Chapter 5. I will assess direct viral infection through alignment against the influenza genome components, and by looking for increased expression of intracellular machinery commonly high-jacked by replicating viruses. I plan to perform the analysis of these data upon successful submission of my dissertation.

We have high expectation for these results. Not only could infection-induced peripheral blood plasmablasts reveal mucosal homing transcriptional profiles, but they could also provide insight on why infection-induced protection is in general longer lasting and offers broader protection (Wang et al., 2011; Yu et al., 2008). It is very likely that the inflammatory environment present during plasmablast differentiation in the two conditions is significantly different, and this may be evident in the plasmablast transcriptome. For example, it is possible the infection-induced environment has improved capacity to induce long lived plasma cell (LLPC) differentiation, which could be responsible for the increased duration of protection. Cells destined to become LLPC could be identified by increased expression of CD138, CXCR4, CD93, CD28, CD44, VLA4, IL-6R, BCMA, MCL1, ZBTB20 in addition to many other genes (Nutt et al., 2015). It is also possible that even without increased LLPC differentiation, the much larger immune response orchestrated by infection delivers additional signals (perhaps from recruited innate cells) that simply generate better plasmablasts (Wrammert et al., 2012).

These studies will further our understanding of the differences between immune responses to infection and vaccination which is already an important goal for the field of vaccine immunology.

Identification of cells orchestrating a broader immune response will eventually require careful characterization of the antibody produced by each cell. A broadly protective antibody is one that is capable of binding to, and neutralizing, antigenically diverse influenza strains. These antibodies target specific highly conserved epitopes on the surface proteins, HA and NA (Neu et al., 2016). Some broadly protective antibodies also rely on poly-reactivity (Andrews et al., 2015) and therefore this feature should also be explored. It is unclear if the transcriptome will reflect this level of specificity detail, but identification of a poly-reactive B cell transcriptome would have significant impact for the field of autoimmunity and therefore should be assessed.

It is currently unclear if influenza viruses can infect human peripheral blood B cells. It is also unknown if BCR specificity impacts B cell infection. We hypothesize, that a virus specific BCR could enhance viral infection of B cells either by increasing membrane contact or through the pathway of receptor mediated endocytosis. With Spec-seq we can identify the influenza virus reactive or non-reactive plasmablasts and directly assess viral infection giving insight into this unresolved question. Furthermore, it seems logical that these B cells may initiate pathways to prevent viral infection as viral infection of your protective antibody secreting cells could impair the overall humoral immune response. We will therefore also look for evidence of protective machinery preferentially up-regulated in infection versus vaccination settings.

This study will be the first ever comparison of the single cell gene expression profiles of B cells during an ongoing influenza infection versus a vaccine response. Therefore, regardless of what the results show, they will give key insight into how similar our seasonal influenza vaccine is to naturally occurring infection and should help guide improved vaccine design.

4.3 Applications for Spec-seq

As mentioned in Chapter 2, Spec-seq is already being applied to numerous other questions within, and outside of, the Wilson Lab. We believe this unique combination of specificity and transcriptome data is ideal to further our understanding of the origin of antigen-specific cells at different time points post immunization, as well as the characterization of B cells outside of the peripheral blood compartment.

In addition, our success applying this scRNA-seq protocol to samples containing 20-100 cells is extremely sought after. This is because problems have been encountered with transcriptional profiling of samples between 1 and 3000 cells. Our consideration of the 100-cell wells when optimizing this protocol was intended as a population level control for understanding transcriptional heterogeneity. However, this has unintentionally facilitated transcriptional analysis of much smaller bulk populations. We have also experienced success applying this approach to low concentrations of RNA extracted from samples containing 800-10,000 cells (Lau et al., 2017).

5. Experimental procedures

Human peripheral blood isolation

15 healthy subjects were recruited to receive the 2014-15 quadrivalent influenza vaccine (QIV) Fluarix© (GSK) at the University of Chicago, with approval by the Institutional Review Board (Protocol: 09-043A). Forty to eighty milliliters of peripheral blood were drawn seven days post immunization. B cells were enriched from the peripheral blood mononuclear cells (PBMCs) by RosetteSep Human B cell Enrichment kit (Stemcells technologies, cat.15024) treatment and isolated with a Lymphoprep gradient (Corning, cat. 25-072). The six subjects with the highest vaccine positive antibody-secreting cell frequency determined by ELISPSOT were selected for inclusion in our study.

ELISPOT

B cell enriched PBMCs (.25-1e6 cells) were transferred onto ELISPOT plates (Millipore, cat. MSHAN4B50). Plates were coated with goat anti-human immunoglobulin (whole IgG, IgM and IgA, KPL, cat. 01-10-07) at 5µg/ml or with the total administered quadrivalent influenza vaccine at 10µl/ml both in PBS overnight at 4°C and then blocked with complete media for 2 hours at 37°C before the cells were transferred. Plates with cells were incubated overnight at 37°C. The plates were washed and then incubated with biotinylated IgG (Southern Biotech, cat. 2040-08) and biotinylated IgA (Southern Biotech, cat. 2050-08) in separate columns, for each coated antigen type for 2.5 hours at RT. After washing, streptavidin-Alkaline Phosphatase (Southern Biotech, cat. 7100-

04) was added for 2.5 hours at RT and NBT/BCIP (Thermo Scientific, cat. 34032) was added to each well for spot reveal.

Cell sorting

The remaining enriched peripheral blood B cells were stained for Fluorescence-Activated-Cell-Sorting (FACS) on either a BD FACS Aria II or the BD FACS Aria Fusion at the University of Chicago flow cytometry core. Antibodies used included: CD19 (BioLegend cat. 30223, clone H1B19), CD27 (BioLegend cat. 302829, clone O323), CD38 (BioLegend cat. 303515, clone HIT2), IgG (BD cat. 550931, clone G18-145), IgA (Southern Biotech cat. 2050-02, clone G1012-YG02B) and IgM (Southern Biotech cat. 9022-09, clone UHB). After gating on lymphocytes, doublet exclusion was performed. The CD19+ B cells were visualized to estimate the total CD27++ CD38++ plasmablast (PB) response. Samples were required to have more than a 0.5% PB response for inclusion in the study. The IgM- CD19+ CD27++ and CD38++ PBs were bulk sorted by IgG and IgA surface expression, prior to single-cell sorting into 96 well plates (BioRad) containing 4µl of the catching solution described previously (Picelli et al., 2014). Plates were immediately sealed, placed on dry ice and frozen at -80°C until use. Each plate contained 41 single cells, 6 unsorted wells and one well with 100 cells.

Simultaneous single cell RNA sequencing and monoclonal antibody generation

This section outlines the modifications required to merge the two pre-existing protocols as well our optimized modifications to the SmartSeq2 protocol (Picelli et al., 2014).

Here we would also like to draw attention to a few improvements that were made between the processing of the first two tagmentation plates, and the later 3 plates. PW2 and PW3 were processed with the SuperScript III reverse transcriptase (RT) enzyme. Whereas, PW4-6 were processed with the Primescript RT enzyme and 5'Biotin modifications were added to all three oligomers; Oligo-DT, TSO and IS-PCR. We believe all effects of these modifications on the transcriptome data have been removed with experimental batch correction.

Cell isolation:

Plates for single cell sorting are prepared in an RNA-only hood, with pipettes used exclusively in that hood, filtered pipette tips, and the user wears a disposable sterile gown with two sets of gloves. The final set of gloves needs to be donned immediately before preparing the mixture, and need to be replaced if they touch any unclean object. Prior to use the hood, racks and pipettes are always cleaned with DNAoff (Takara Bio 9036) and RNaseZap (Invitrogen AM9788). Everything that is brought into the hood is also cleaned.

The catching solution used for single cell sorting was identical to that described in the SmartSeq2 protocol (Picelli et al., 2014), except for the oligomer biotin modification mentioned above to reduce undesired primer to primer annealing and amplification. FACS was used to isolate the cells of interest and every sorted plate contained 41 single cells, 6 wells with no cell, and a single 100-cell well. The 100-cell wells were processed at the same time and with the same protocol as used for the single-cell wells.

cDNA reaction and amplification:

The plate to be processed was thawed on ice while the reverse transcription (RT) mixture was made in the RNA-only hood. For Primescript reactions the RT-mixture contains 100 Units of Primescript (Clonetech, cat. 2680), 10 Units RNase inhibitor (Clonetech, cat. 2313), 1M Betaine (Sigma Aldrich, cat. B0300), 6mM $MgCl_2$ (Life technologies, cat. AM9530G) and 1μM 5' Biotin - TSO oligomer (IDT). The reaction volume was raised to 5.7μl with nuclease free water. When the SuperScript III enzyme (Invitrogen, cat. 180800) was used, 5mM DTT (Invitrogen, cat. 180800) was also included, as suggested in the SmartSeq2 protocol. The Oligo-DT reaction (72°C for 3 minutes, with a heated lid at 80°C, then hold at 4°C) was run prior to the addition of the RT mixture. Then the SmartSeq2 RT PCR reaction was used for all plates regardless of the RT enzyme. SmartSeq2 RT PCR reaction: 42°C for 90 minutes, 10 cycles of 50°C for 2 minutes and 42°C for 2 minutes, then 70°C for 15 minutes before a 4°C hold. The heated lid was at 80°C.

Once complete the pre-amplification PCR mix was assembled on a sterile bench top as described previously (Picelli et al., 2014). 22 pre-amplification PCR cycles was selected for our PB populations of interest, and amplification was evaluated on an Agilent high-sensitivity DNA chip (cat. 5067-4626) (Fig. 2.1b, Fig. 2.3). All PB traces obtained during the course of this experiment contained two sharp amplicon peaks suggestive of extremely high expression for certain transcripts, the smaller of which is likely the BCR itself. The pre-amplification PCR was as follows: Denature at 98°C for 3 minutes, then 22 cycles of 98°C for 20 seconds, 67°C for 15 seconds and 72°C for 6 minutes, followed by 72°C for 5 minutes before a hold at 4°C.

After bead based PCR purification (Picelli et al., 2014) 1µl of cDNA from each cell was used in three separate PCRs designed to amplify the heavy, kappa and lambda chain of the BCR. The remaining cDNA was stored at -80°C until next generation sequencing library preparation.

Antibody PCR amplification and cloning:

The antibody cloning of samples collected in 2014-15 (PW2 & PW3) was performed as described in Smith et al., 2009 (Smith et al., 2009) whereas the samples collected in 2015-16 (PW4-6) followed the revised protocol described in Ho et al., 2016 (Ho et al., 2016). No differences in antibody binding were observed with the two different cloning approaches (Ho et al., 2016).

Briefly, both approaches utilized separate sets of PCR to amplify the heavy chain, the kappa chain and the lambda chain. The 1st PCR used a cocktail of primers designed to bind to all possible V gene and constant domains utilized by the respective antibody chain. The cocktail of primers used in the 2nd PCR are nested inside the first primer binding sites to allow for additional amplification of the generated PCR product. Then Sanger sequencing identified the specific V(D)J or VJ combination used in each antibody chain. This information guided the primers for the cloning PCR, which incorporates either the necessary restriction digest sites for vector ligation (Smith et al., 2009) or the required sequence overhang for Gibson Assembly (Ho et al., 2016). Regardless of the cloning method used, each antibody heavy chain V(D)J was inserted into the same IgG1 backbone within the AbVec expression vector, whereas each light chain VJ portion was inserted into their respective kappa or lambda AbVec expression

vector. All details of this approach are clearly outlined in the referenced papers (Ho et al., 2016; Smith et al., 2009).

Library generation, pooling & sequencing:

The protocol for library generation proceeds according to the SmartSeq2 protocol (Picelli et al., 2014), which we will cover briefly in this section.

Samples with acceptable amplification peaks then underwent library generation with the Nextera XT DNA sample prep kit (Illumina FC-131-1096). The tagmentation is first, and this step will break the cDNA into small fragments with specific adaptors being attached at the 5' and 3' ends of each fragment, which allows for subsequent index incorporation. This reaction uses 10µl of the Tagmentation DNA buffer, 5µl of the Amplicon tagment mix and [.5 ng] of the amplified cDNA. Then the tagmentation PCR reaction was run; 55°C for 5 minutes before a 4°C hold.

Immediately after the PCR reaction was complete 5µl of the neutralizing NT buffer was added to every well, mixed, and incubated for 5 minutes at room temperature. Next the Nextera indexes are added to every fragment through PCR amplification with the Nextera XT Index Kit (Illumina FC-131-1002). This reaction is performed directly on the tagmented DNA and uses 15µl of the Nextera PCR master mix, 5µl of an Index-1 primer and 5µl of an Index-2 primer. There are 8 Index-1 primers where each one is used across an entire row of samples and then the 12 Index-2 primers are used one per column. This results in a unique combination of indexes for each library included in the 96-well plate. With our samples we decided to perform 12 PCR cycles. This PCR reaction is: 72°C for 3 minutes, 95°C for 30 seconds, 12 cycles

of 95°C for 10 seconds, 55°C for 30 second and 72°C for 30 seconds were performed, before 5 minutes at 72°C and a 4°C hold.

These samples are then PCR purified with the Ampure XP beads again prior to Agilent high sensitivity DNA chip analysis. This protocol is the same as before, except the recommended ratio of sample volume to beads is now 3:5. This should help reduce carryover of the smaller non-specific fragments.

The Agilent bioanalyzer high-sensitivity DNA chip was used to measure fragment size and concentration. Equal volumes of all libraries were then pooled into a single tube at 3 nM each and underwent paired-end 50 base pair sequencing in two lanes on the Illumina HiSeq2500 in the University of Chicago's Genomics Facility.

Antibody vector transfection

Antibody vector pairs (heavy and light chains) were then transfected into 293-A cells (Invitrogen, cat. R70507) for human monoclonal antibody production, as described previously (Smith et al., 2009). This protocol was modified slightly to allow for small volume, more high-throughput transfections. This involved coating 24 well tissue culture plates with $2.5x10^5$ 293-A cells in 1 ml cell media 12-18 hours before transfection. For transfection, 360 ng of each vector are mixed in 200 µl plain DMEM (Invitrogen, cat. 11995081) with 4µl polyethylenimine (Polysciences, cat. 23966-1) at 1mg/ml, incubated at room temperature for 15 minutes prior to addition to the appropriate well. During the incubation period 500 µl of media was removed from each plate well. After 8-16 hours the transfection mixture was removed and replaced with PFHMII-Protein Free Hybridoma media (Invitrogen, cat. 12040-077). Four to five days later the supernatant

was collected, centrifuged for ten minutes at 3000 rpm to remove cell debris, and then used directly in screening ELISAs.

Screening ELISA

To quantify supernatant IgG content, ELISA plates (Fisher, cat. 3369) were coated with anti-human IgG (Jackson Immuno Research, cat. 109-1006-098) at 2 μg/ml in carbonate buffer. To quantify antigen binding, plates were coated with a 1:100 dilution of the influenza vaccine (Fluarix© QIV, GSK) in PBS. After 12-24 hours at 4°C, the plates were washed and blocked with 20% FBS for 1 hour at 37°C. A 1:25 supernatant dilution was the starting concentration for the anti-IgG plates, whereas the supernatant was used at full concentration on the influenza antigen plates. In both situations the starting solution undergoes serial dilutions down the plate to obtain a binding curve. For the anti-IgG plate, the positive control was purified human IgG (Thermo Scientific, cat. 31154) at a concentration of 250 ng/ml. For the influenza vaccine-coated plates, the well-characterized CR9114 antibody (Dreyfus et al., 2012) was used at a concentration of 10μg/ml. A horseradish peroxidase-conjugated anti-human IgG Fc-specific antibody (Mabtech, cat. 3820-4-250) was used for detection and the plates were developed with Super Aquablue ELISA Substrate (eBiosciences, cat. 00-4203-58). Absorbance was measured at a wavelength of 405 nm on a microplate spectrophotometer (BioRad). If a sample had an OD (minus background) greater than 0.5 on the anti-IgG plate once the positive control has reached 1.0, it was considered a successful transfection. The samples on the influenza plates must also have an OD greater than 0.5 when the positive control OD reaches 3.0. The requirement for an antigen binding OD greater

than 0.5 is based on previous work in our lab where 0.5 was identified as two standard deviations above the mean absorbance of antibodies form naive B cells (Wrammert et al., 2011). Experimental replicates were performed for each antibody to confirm the results.

scRNA-seq data analysis

Pipeline overview in Figure 3.3a

Step 1: Repertoire

Sanger BCR sequences were obtained from the nested PCR and the cloning PCR. Assembled BCR contigs were generated by applying the BASIC (Canzar et al., 2016) algorithm to the raw transcriptome files for all single cells. The BCR sequences from PW2 and PW3 obtained by all three approaches were carefully compared previously (Canzar et al., 2016), to confirm the accuracy of BASIC. In general the assembled sequences provided by BASIC had less error and often reported a slightly reduced mutation count. This could in part be due to the high number of PCR cycles required to amplify the BCR. Therefore, we consider the assembly by BASIC to be a more accurate report of the original BCR sequence and used that contig, when available, for repertoire analysis (Table 3.1-3.2). If BASIC was unable to assemble a full-length sequence the Sanger sequence obtained from the cloning PCR was used in its place.

Clones were identified through the same usage of V and J genes, as well as highly similar CDR3 sequences. When searching for any repertoire differences between the three populations of interest (Fig. 3.5), all clonal expansions were excluded, as their

inclusion could be confused with population gene preferences. However, when validating the identified DEGs clonal expansions were not eliminated, as those cells were present in the transcriptome (Fig. 3.8f).

Step 2: Pseudoalignment and transcript quantification

Kallisto (Bray et al., 2016b; Ntranos et al., 2016) was used for pseudoalignment of the raw sequencing data against the human transcriptome. The *-quant* command (with the *--bias* option) was used to perform pseudo-alignment of each samples data to the human transcriptome. Expression quantification is reported as Transcripts Per Million, which is adjusted for transcript length and library size, and facilitates direct comparison of gene expression profiles across cells.

Step 3: gene expression dataframe, transformation, quality filters & gene filters

The dataframe of gene expression was summarized by summing each genes transcript variants prior to undergoing Log2(TPM+1) transformation. Next all values below 0.4, our empirically identified threshold of detection (Fig. 3.3b), were set to 0 and all genes whose expression was never above 0 were removed from the dataset. Our first cell quality filter excluded cells where less than 40% of the total reads mapped to the human transcriptome (Fig. 3.3c). Our next cell filter excluded cells where the V, J and Constant domains did not match between at least two of our three approaches; Sanger sequencing of the nested PCR, of the Cloning PCR and the assembled BASIC sequence. This filer simultaneously confirms that each sample only contains a single cell. The constant domain of each BCR sequence must also match the flow cytometry identified isotype. Scatter plots between single cells reveal significant variability in pairwise Pearson correlation coefficients (Fig. 3.3d & e), suggesting a high level of

population heterogeneity. However, when the single cells are averaged and plotted against a 100-cell well sample (Fig. 3.3f) the correlation coefficient matches that obtained between two 100-cell wells (Fig. 3.3g). Therefore, the majority of the heterogeneity is likely biological, versus technical, in nature. Next, a gene filter was applied that requires expression in at least 5% of the cells, reducing the dataframe to 11895 genes.

Step 4: quantile normalization and batch correction

Quantile normalization using the preprocesscore package (Bolstad et al., 2003) was performed within each experimental plate to reduce differences between the quantiles of cell-specific TPM counts. Next, the Limma package (Ritchie et al., 2015) was used to remove known batch effects associated with experimental plate and patient ID, while protecting our variables of interest; BCR isotype and vaccine recognition (Fig. 3.4):

```
adj_DF <- removeBatchEffect (QN_DF, batch= Experiment, batch2 = Patient ID, design = model.matrix(~Isotype + Vaccine_binding))
```

Step 5: Differential gene identification & visualization

Differentially expressed genes were identified with a t-test, where samples were either grouped by BCR isotype or BCR vaccine reactivity. We required a fold change greater than or equal to 1.5 and a Benjamini-Hochberg adjusted p-value less than 0.05. Visualization of the scRNA-seq data was obtained with the Rtsne package (Krijthe, 2017). Default parameters were used except theta = 0.001 and initial dims = 10. The same parameters were used for all datasets.

Statistical analysis

A paired t-test was used to assess significant differences between ELISPOT and mAb vaccine binding (Fig. 3.1f) and the cell cycle phase assignments (Fig. 3.8e & Fig. 3.11e). One-way analysis of variance (ANOVA) was used to identify statistically significant differences (Fig. 3.5 e-h). To assess global transcriptional similarity within vaccine positive and negative cells we performed two quantitative steps. First, we calculated the Pearson correlation between cells within vaccine positive clones, and between vaccine negative cells. Second, since correlation coefficients are bounded in [-1, +1], we performed a Fisher's Z transformation [arctanh(p)] on the coefficients and subjected the transformed values to a one-sided two-sample t-test. Specifically, we tested if the average correlation between clonally related cells is significantly greater than non-clonally related cells in various instances.

Analysis scripts

The R script used to analyze the scRNA-seq data is contained in Figure 5.1.

6. References

Andrews, S.F., Huang, Y., Kaur, K., Popova, L.I., Ho, I.Y., Pauli, N.T., Dunand, C.J.H., Taylor, W.M., Lim, S., Huang, M., et al. (2015). Immune history profoundly affects broadly protective B cell responses to influenza. Science Translational Medicine *7*, 316ra192–ra316ra192.

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biology *17*, 63.

Barría, M.I., Garrido, J.L., Stein, C., Scher, E., Ge, Y., Engel, S.M., Kraus, T.A., Banach, D., and Moran, T.M. (2013). Localized Mucosal Response to Intranasal Live Attenuated Influenza Vaccine in Adults. J Infect Dis *207*, 115–124.

Basset, Devauchelle, Durand, Jamin, Pennec, Youinou, and Dueymes (1999). Glycosylation of Immunoglobulin A Influences Its Receptor Binding. Scandinavian Journal of Immunology *50*, 572–579.

Beck, A., Goetsch, L., Dumontet, C., and Corvaïa, N. (2017). Strategies and challenges for the next generation of antibody-drug conjugates. Nat Rev Drug Discov *16*, 315–337.

Beeler, J.A., and van Wyke Coelingh, K. (1989). Neutralization epitopes of the F glycoprotein of respiratory syncytial virus: effect of mutation upon fusion function. J Virol *63*, 2941–2950.

Von Behring, E., and Kitasato, S. (1890). [The mechanism of diphtheria immunity and tetanus immunity in animals. 1890]. Mol. Immunol. *28*, 1317, 1319–1320.

Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. Cell *157*, 714–725.

Berton, M.T., Uhr, J.W., and Vitetta, E.S. (1989). Synthesis of germ-line gamma 1 immunoglobulin heavy-chain transcripts in resting B cells: induction by interleukin 4 and inhibition by interferon gamma. Proc Natl Acad Sci U S A *86*, 2829–2833.

Björklund, Å.K., Forkel, M., Picelli, S., Konya, V., Theorell, J., Friberg, D., Sandberg, R., and Mjösberg, J. (2016). The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. Nat Immunol *17*, 451–460.

Bolstad, B.M., Irizarry, R.A., Åstrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Boyd, S.D., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., Simen, B.B., Hanczaruk, B., Nguyen, K.D., et al. (2009). Measurement and

Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. Science Translational Medicine *1*, 12ra23–12ra23.

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

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

Brüggemann, M., Williams, G.T., Bindon, C.I., Clark, M.R., Walker, M.R., Jefferis, R., Waldmann, H., and Neuberger, M.S. (1987). Comparison of the effector functions of human immunoglobulins using a matched set of chimeric antibodies. J. Exp. Med. *166*, 1351–1361.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotech *33*, 155–160.

Bullough, P.A., Hughson, F.M., Skehel, J.J., and Wiley, D.C. (1994). Structure of influenza haemagglutinin at the pH of membrane fusion. Nature *371*, 37–43.

Bunker, J.J., Erickson, S.A., Flynn, T.M., Henry, C., Koval, J.C., Meisel, M., Jabri, B., Antonopoulos, D.A., Wilson, P.C., and Bendelac, A. (2017). Natural polyreactive IgA antibodies coat the intestinal microbiota. Science eaan6619.

Burnet, F.M. (1976). A modification of Jerne's theory of antibody production using the concept of clonal selection. CA Cancer J Clin *26*, 119–121.

Butcher, E.C., Rouse, R.V., Coffman, R.L., Nottenburg, C.N., Hardy, R.R., and Weissman, I.L. (1982). Surface phenotype of Peyer's patch germinal center cells: implications for the role of germinal centers in B cell differentiation. J. Immunol. *129*, 2698–2707.

Cann, G.M., Gulzar, Z.G., Cooper, S., Li, R., Luo, S., Tat, M., Stuart, S., Schroth, G., Srinivas, S., Ronaghi, M., et al. (2012). mRNA-Seq of Single Prostate Cancer Circulating Tumor Cells Reveals Recapitulation of Gene Expression and Pathways Found in Prostate Cancer. PLOS ONE *7*, e49144.

Canzar, S., Neu, K.E., Tang, Q., Wilson, P.C., and Khan, A.A. (2016). BASIC: BCR assembly from single cells. Bioinformatics btw631.

CDC (2017). Seasonal Influenza Vaccine Effectiveness, 2005-2017. https://www.cdc.gov/flu/professionals/vaccination/effectiveness-studies.htm

Chan, A.C., and Carter, P.J. (2010). Therapeutic antibodies for autoimmunity and inflammation. Nat Rev Immunol *10*, 301–316.

Chaussabel, D., Pascual, V., and Banchereau, J. (2010). Assessing the human immune system through blood transcriptomics. BMC Biology *8*, 84.

Chen, K., and Cerutti, A. (2010). Vaccination strategies to promote mucosal antibody responses. Immunity *33*, 479–491.

Chen-Kiang, S. (2003). Cell-cycle control of plasma cell differentiation and tumorigenesis. Immunological Reviews *194*, 39–47.

Chuang, H.-Y., Hofree, M., and Ideker, T. (2010). A Decade of Systems Biology. Annu Rev Cell Dev Biol *26*, 721–744.

Dreyfus, C., Laursen, N.S., Kwaks, T., Zuijdgeest, D., Khayat, R., Ekiert, D.C., Lee, J.H., Metlagel, Z., Bujny, M.V., Jongeneelen, M., et al. (2012). Highly conserved protective epitopes on influenza B viruses. Science *337*, 1343–1348.

Dueck, H., Khaladkar, M., Kim, T.K., Spaethling, J.M., Francis, C., Suresh, S., Fisher, S.A., Seale, P., Beck, S.G., Bartfai, T., et al. (2015). Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. Genome Biology *16*, 122.

Edelman, G. (1972). ANTIBODY STRUCTURE AND MOLECULAR IMMUNOLOGY. Nobel Lecture.

Ehrlich, P. (1902). The side chain theory of immunity. Instructions for hygienic examinations: according to the principles laid down in the Hygienic Institute of the Royal Order. Ludwig-Maximilians-University of Munich, 3 editions 3.

Eisen, H.N., and Siskind, G.W. (1964). VARIATIONS IN AFFINITIES OF ANTIBODIES DURING THE IMMUNE RESPONSE. Biochemistry *3*, 996–1008.

Eltahla, A.A., Rizzetto, S., Rasoli, M., Betz-Stablein, B.D., Venturi, V., Kedzierska, K., Lloyd, A.R., Bull, R.A., and Luciani, F. (2016). Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. Immunol Cell Biol.

Emmons, C., and Hunsicker, L.G. (1987). Muromonab-CD3 (Orthoclone OKT3): the first monoclonal antibody approved for therapeutic use. Iowa Med *77*, 78–82.

Ezhevsky, S.A., Toyoshima, H., Hunter, T., and Scott, D.W. (1996). Role of cyclin A and p27 in anti-IgM induced G1 growth arrest of murine B-cell lymphomas. Mol. Biol. Cell *7*, 553–564.

Fan, H.C., Fu, G.K., and Fodor, S.P.A. (2015). Combinatorial labeling of single cells for gene expression cytometry. Science *347*, 1258367.

FDA (2017). New FDA Approved Drugs for 2017 | CenterWatch. http://www.centerwatch.com/drug-information/fda-approved-drugs/

Fulwyler, M.J. (1965). Electronic Separation of Biological Cells by Volume. Science *150*, 910–911.

Gaublomme, J.T., Yosef, N., Lee, Y., Gertner, R.S., Yang, L.V., Wu, C., Pandolfi, P.P., Mak, T., Satija, R., Shalek, A.K., et al. (2015). Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. Cell *163*, 1400–1412.

Ge, X., Mazor, Y., Hunicke-Smith, S.P., Ellington, A.D., and Georgiou, G. (2010). Rapid construction and characterization of synthetic antibody libraries without DNA amplification. Biotechnol. Bioeng. *106*, 347–357.

Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotech *32*, 158–168.

Gilad, Y., and Mizrahi-Man, O. (2015). A reanalysis of mouse ENCODE comparative gene expression data. F1000Research.

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G.R., Ni, I., Mei, L., Sundar, P.D., Day, G.M.R., et al. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A *106*, 20216–20221.

Gomes, M.M., Wall, S.B., Takahashi, K., Novak, J., Renfrow, M.B., and Herr, A.B. (2008). Analysis of IgA1 N-glycosylation and its contribution to FcαRI binding. Biochemistry *47*, 11285–11299.

Gonzalez-Quintela, A., Alende, R., Gude, F., Campos, J., Rey, J., Meijide, L.M., Fernandez-Merino, C., and Vidal, C. (2008). Serum levels of immunoglobulins (IgG, IgA, IgM) in a general adult population and their relationship with alcohol consumption, smoking and common metabolic abnormalities. Clin Exp Immunol *151*, 42–50.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech *29*, 644–652.

Gray, J.W., Carrano, A.V., Steinmetz, L.L., Dilla, M.A.V., Moore, D.H., Mayall, B.H., and Mendelsohn, M.L. (1975). Chromosome measurement and sorting by flow systems. PNAS *72*, 1231–1234.

Grillo-López, A.J., White, C.A., Varns, C., Shen, D., Wei, A., McClure, A., and Dallaire, B.K. (1999). Overview of the clinical development of rituximab: first monoclonal antibody approved for the treatment of lymphoma. Semin. Oncol. *26*, 66–73.

Grün, D., and van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. Cell *163*, 799–810.

Hayes, J.M., Cosgrave, E.F.J., Struwe, W.B., Wormald, M., Davey, G.P., Jefferis, R., and Rudd, P.M. (2014). Glycosylation and Fc receptors. Curr. Top. Microbiol. Immunol. *382*, 165–199.

Hicks, S.C., Teng, M., and Irizarry, R.A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. bioRxiv 025528.

Ho, I.Y., Bunker, J.J., Erickson, S.A., Neu, K.E., Huang, M., Cortese, M., Pulendran, B., and Wilson, P.C. (2016). Refined protocol for generating monoclonal antibodies from single human and murine B cells. Journal of Immunological Methods.

Hodoniczky, J., Zheng, Y.Z., and James, D.C. (2005). Control of recombinant monoclonal antibody effector functions by Fc N-glycan remodeling in vitro. Biotechnol. Prog. *21*, 1644–1652.

Horns, F., Vollmers, C., Croote, D., Mackey, S.F., Swan, G.E., Dekker, C.L., Davis, M.M., and Quake, S.R. (2016). Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. eLife *5*, e16578.

Ichinohe, T., Tamura, S., Kawaguchi, A., Ninomiya, A., Imai, M., Itamura, S., Odagiri, T., Tashiro, M., Takahashi, H., Sawa, H., et al. (2007). Cross-Protection against H5N1 Influenza Virus Infection Is Afforded by Intranasal Inoculation with Seasonal Trivalent Inactivated Influenza Vaccine. J Infect Dis. *196*, 1313–1320.

Islam, K.B., Nilsson, L., Sideras, P., Hammarström, L., and Smith, C.I. (1991). TGF-beta 1 induces germ-line transcripts of both IgA subclasses in human B lymphocytes. Int. Immunol. *3*, 1099–1106.

Iversen, R., Snir, O., Stensland, M., Kroll, J.E., Steinsbø, Ø., Korponay-Szabó, I.R., Lundin, K.E.A., Souza, G.A. de, and Sollid, L.M. (2017). Strong Clonal Relatedness between Serum and Gut IgA despite Different Plasma Cell Origins. Cell Reports *20*, 2357–2367.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science *343*, 776–779.

Jenner, E. (1798). An Inquiry Into the Causes and Effects of the Variole Vaccine, or Cow-Pox.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat *8*, 118–127.

Kataoka, T., Kawakami, T., Takahashi, N., and Honjo, T. (1980). Rearrangement of immunoglobulin gamma 1-chain gene and mechanism for heavy-chain class switch. Proc Natl Acad Sci U S A *77*, 919–923.

Kavaler, J., Davis, M.M., and Chien, Y. (1984). Localization of a T-cell receptor diversity-region element. Nature *310*, 421–423.

Keimpema, M. van, Grüneberg, L.J., Mokry, M., Boxtel, R. van, Zelm, M.C. van, Coffer, P., Pals, S.T., and Spaargaren, M. (2015). The forkhead transcription factor FOXP1 represses human plasma cell differentiation. Blood *126*, 2098–2109.

Kerr, M.A. (1990). The structure and function of human IgA. Biochem. J. *271*, 285–296.

Keusch, J., Lydyard, P.M., and Delves, P.J. (1998). The effect on IgG glycosylation of altering beta1, 4-galactosyltransferase-1 activity in B cells. Glycobiology *8*, 1215–1220.

Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nat Meth *11*, 740–742.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods *9*, 72–74.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell *161*, 1187–1201.

Klein, U., Rajewsky, K., and Küppers, R. (1998). Human Immunoglobulin (Ig)M+IgD+ Peripheral Blood B Cells Expressing the CD27 Cell Surface Antigen Carry Somatically Mutated Variable Region Genes: CD27 as a General Marker for Somatically Mutated (Memory) B Cells. J Exp Med *188*, 1679–1689.

Köhler, G., and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. Nature *256*, 495–497.

Krammer, F. (2015). The Quest for a Universal Flu Vaccine: Headless HA 2.0. Cell Host & Microbe *18*, 395–397.

Krijthe, J. (2017). Rtsne: R wrapper for Van der Maaten's Barnes-Hut implementation of t-Distributed Stochastic Neighbor Embedding.

Kuraoka, Masayuki (2016). B-Cell Selection in Germinal Centers Elicited by Complex Antigens. Immunity.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.

Lau, D., Lan, L.Y.-L., Andrews, S.F., Henry, C., Rojas, K.T., Neu, K.E., Huang, M., Huang, Y., DeKosky, B., Palm, A.-K.E., et al. (2017). Low CD21 expression defines a population of recent germinal center graduates primed for plasma cell differentiation. Science Immunology *2*, eaai8153.

Laver, W.G., and Valentine, R.C. (1969). Morphology of the isolated hemagglutinin and neuraminidase subunits of influenza virus. Virology *38*, 105–119.

Lawton, A.R., Kincade, P.W., and Cooper, M.D. (1975). Sequential expression of germ line genes in development of immunoglobulin class diversity. Fed. Proc. *34*, 33–39.

Leek, J.T., and Storey, J.D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLOS Genet *3*, e161.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet *11*, 733–739.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500.

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., et al. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat Immunol *15*, 195–204.

Liao, H.-X., Lynch, R., Zhou, T., Gao, F., Alam, S.M., Boyd, S.D., Fire, A.Z., Roskin, K.M., Schramm, C.A., Zhang, Z., et al. (2013). Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature *496*, 469–476.

Lindhout, E., Mevissen, M.L., Kwekkeboom, J., Tager, J.M., and de Groot, C. (1993). Direct evidence that human follicular dendritic cells (FDC) rescue germinal centre B cells from death by apoptosis. Clin. Exp. Immunol. *91*, 330–336.

Littman, D.R. (2015). Releasing the Brakes on Cancer Immunotherapy. Cell *162*, 1186–1190.

Lundgren, M., Persson, U., Larsson, P., Magnusson, C., Smith, C.I., Hammarström, L., and Severinson, E. (1989). Interleukin 4 induces synthesis of IgE and IgG4 in human B cells. Eur. J. Immunol. *19*, 1311–1315.

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research *9*, 2579–2605.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Mahata, B., Zhang, X., Kolodziejczyk, A.A., Proserpio, V., Haim-Vilmovsky, L., Taylor, A.E., Hebenstreit, D., Dingler, F.A., Moignard, V., Göttgens, B., et al. (2014). Single-Cell RNA Sequencing Reveals T Helper Cells Synthesizing Steroids De Novo to Contribute to Immune Homeostasis. Cell Reports *7*, 1130–1142.

Maverakis, E., Kim, K., Shimoda, M., Gershwin, M.E., Patel, F., Wilken, R., Raychaudhuri, S., Ruhaak, L.R., and Lebrilla, C.B. (2015). Glycans in the immune system and The Altered Glycan Theory of Autoimmunity: a critical review. J. Autoimmun. *57*, 1–13.

McKean, D., Huppi, K., Bell, M., Staudt, L., Gerhard, W., and Weigert, M. (1984). Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. Proc. Natl. Acad. Sci. U.S.A. *81*, 3180–3184.

Medina, R.A., and García-Sastre, A. (2011). Influenza A viruses: new research developments. Nat. Rev. Microbiol. *9*, 590–603.

Mei, H.E., Yoshida, T., Sime, W., Hiepe, F., Thiele, K., Manz, R.A., Radbruch, A., and Dörner, T. (2009). Blood-borne human plasma cells in steady state are derived from mucosal immune responses. Blood *113*, 2461–2469.

Meredith, M., Zemmour, D., Mathis, D., and Benoist, C. (2015). Aire controls gene expression in the thymic epithelium with ordered stochasticity. Nat Immunol *16*, 942–949.

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

Murphy, K., and Weaver, C. (2016). Janeway's Immunobiology, 9th edition (Garland Science).

Nakaya, H.I., and Pulendran, B. (2012). Systems vaccinology: its promise and challenge for HIV vaccine development. Curr Opin HIV AIDS *7*, 24–31.

Nakaya, H.I., Wrammert, J., Lee, E.K., Racioppi, L., Marie-Kunze, S., Haining, W.N., Means, A.R., Kasturi, S.P., Khan, N., Li, G.-M., et al. (2011). Systems biology of vaccination for seasonal influenza in humans. Nat. Immunol. *12*, 786–795.

Neu, K.E., Henry Dunand, C.J., and Wilson, P.C. (2016). Heads, stalks and everything else: how can antibodies eradicate influenza as a human disease? Current Opinion in Immunology *42*, 48–55.

Neutra, M.R., and Kozlowski, P.A. (2006). Mucosal vaccines: the promise and the challenge. Nat Rev Immunol *6*, 148–158.

Nieuwenhuis, P., and Opstelten, D. (1984). Functional anatomy of germinal centers. Am. J. Anat. *170*, 421–435.

Nobelprize (2010). Nobel Prizes and the Immune System.

Nossal, G.J.V., Szenberg, A., Ada, G.L., and Austin, C.M. (1964). Single Cell Studies on 19s Antibody Production. Journal of Experimental Medicine *119*, 485–502.

Ntranos, V., Kamath, G.M., Zhang, J.M., Pachter, L., and Tse, D.N. (2016). Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. Genome Biology *17*, 112.

Nutt, S.L., Hodgkin, P.D., Tarlinton, D.M., and Corcoran, L.M. (2015). The generation of antibody-secreting plasma cells. Nat Rev Immunol *15*, 160–171.

Obermoser, G., Presnell, S., Domico, K., Xu, H., Wang, Y., Anguiano, E., Thompson-Snipes, L., Ranganathan, R., Zeitner, B., Bjork, A., et al. (2013). Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. Immunity *38*, 831–844.

Ohmit, S.E., Petrie, J.G., Cross, R.T., Johnson, E., and Monto, A.S. (2011). Influenza hemagglutination-inhibition antibody titer as a correlate of vaccine-induced protection. J. Infect. Dis. *204*, 1879–1885.

Olsson, A., Venkatasubramanian, M., Chaudhri, V.K., Aronow, B.J., Salomonis, N., Singh, H., and Grimes, H.L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature *537*, 698–702.

Pabst, O. (2012). New concepts in the generation and functions of IgA. Nat Rev Immunol *12*, 821–832.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protocols *9*, 171–181.

Pierson, E., and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biology *16*, 241.

Porter, R.R. (1972). STRUCTURAL STUDIES OF IMMUNOGLOBULINS.

Powell, A.A., Talasaz, A.H., Zhang, H., Coram, M.A., Reddy, A., Deng, G., Telli, M.L., Advani, R.H., Carlson, R.W., Mollick, J.A., et al. (2012). Single Cell Profiling of Circulating Tumor Cells: Transcriptional Heterogeneity and Diversity from Breast Cancer Cell Lines. PLOS ONE *7*, e33788.

Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. (2016). Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data (Proceedings of The 33rd International Conference on Machine Learning).

Pulendran, B., Li, S., and Nakaya, H.I. (2010). Systems Vaccinology. Immunity *33*, 516–529.

Querec, T.D., Akondy, R.S., Lee, E.K., Cao, W., Nakaya, H.I., Teuwen, D., Pirani, A., Gernert, K., Deng, J., Marzolf, B., et al. (2009). Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat. Immunol. *10*, 116–125.

Rabbitts, T.H., Forster, A., Dunnick, W., and Bentley, D.L. (1980). The role of gene deletion in the immunoglobulin heavy chain switch. Nature *283*, 351–356.

Rajewsky, K., Forster, I., and Cumano, A. (1987). Evolutionary and somatic selection of the antibody repertoire in the mouse. Science *238*, 1088–1094.

Ramilo, O., Allman, W., Chung, W., Mejias, A., Ardura, M., Glaser, C., Wittkowski, K.M., Piqueras, B., Banchereau, J., Palucka, A.K., et al. (2007). Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood *109*, 2066–2077.

Ravn, U., Gueneau, F., Baerlocher, L., Osteras, M., Desmurs, M., Malinge, P., Magistrelli, G., Farinelli, L., Kosco-Vilbois, M.H., and Fischer, N. (2010). By-passing in vitro screening--next generation sequencing technologies applied to antibody display and in silico candidate selection. Nucleic Acids Res. *38*, e193.

Reddy, S.T., Ge, X., Miklos, A.E., Hughes, R.A., Kang, S.H., Hoi, K.H., Chrysostomou, C., Hunicke-Smith, S.P., Iverson, B.L., Tucker, P.W., et al. (2010). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotech *28*, 965–969.

Van Riet, E., Ainai, A., Suzuki, T., and Hasegawa, H. (2012). Mucosal IgA responses in influenza virus infections; thoughts for vaccine design. Vaccine *30*, 5893–5900.

Ringnér, M. (2008). What is principal component analysis? Nat Biotech *26*, 303–304.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Rose, M.A. (2014). Mucosal immunization in perspective. Hum Vaccin Immunother *10*, 2115–2117.

Saliba, A.-E., Westermann, A.J., Gorski, S.A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. Nucl. Acids Res. gku555.

Santibanez, T., Kahn, K., Zhai, Y., O'Halloran, A., Liu, L., Bridges, C., Lu, P.-J., Greby, S., Williams, W., and Singleton, J. (2015). Flu Vaccination Coverage, United States,

2015-16 Influenza Season. https://www.cdc.gov/flu/pdf/fluvaxview /2015-16/nfid-coverage-2015-16-final.pdf

Sarkander, J., Hojyo, S., and Tokoyoda, K. (2016). Vaccination to gain humoral immune memory. Clin Transl Immunology *5*, e120.

Schifferli, J.A., Ng, Y.C., and Peters, D.K. (1986). The role of complement and its receptor in the elimination of immune complexes. N. Engl. J. Med. *315*, 488–495.

Selman, M.H.J., Jong, S.E. de, Soonawala, D., Kroon, F.P., Adegnika, A.A., Deelder, A.M., Hokke, C.H., Yazdanbakhsh, M., and Wuhrer, M. (2012). Changes in Antigen-specific IgG1 Fc N-glycosylation Upon Influenza and Tetanus Vaccination. Mol Cell Proteomics *11*, M111.014563.

Seto, J.T., Drzeniek, R., and Rott, R. (1966). Isolation of a low molecular weight sialidase (neuraminidase) from influenza virus. Biochimica et Biophysica Acta (BBA) - Enzymology and Biological Oxidation *113*, 402–404.

Shade, K.-T.C., and Anthony, R.M. (2013). Antibody Glycosylation and Inflammation. Antibodies *2*, 392–414.

Shaffer, A.L., Shapiro-Shelef, M., Iwakoshi, N.N., Lee, A.-H., Qian, S.-B., Zhao, H., Yu, X., Yang, L., Tan, B.K., Rosenwald, A., et al. (2004). XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. Immunity *21*, 81–93.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *advance online publication*.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363–369.

Shields, R.L., Lai, J., Keck, R., O'Connell, L.Y., Hong, K., Meng, Y.G., Weikert, S.H.A., and Presta, L.G. (2002). Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fcgamma RIII and antibody-dependent cellular toxicity. J. Biol. Chem. *277*, 26733–26740.

Skehel, J.J., Bayley, P.M., Brown, E.B., Martin, S.R., Waterfield, M.D., White, J.M., Wilson, I.A., and Wiley, D.C. (1982). Changes in the conformation of influenza virus hemagglutinin at the pH optimum of virus-mediated membrane fusion. Proc. Natl. Acad. Sci. U.S.A. *79*, 968–972.

Smith, K., Garman, L., Wrammert, J., Zheng, N.-Y., Capra, J.D., Ahmed, R., and Wilson, P.C. (2009). Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. Nat. Protocols *4*, 372–384.

Smith, K.G., Hewitson, T.D., Nossal, G.J., and Tarlinton, D.M. (1996). The phenotype and fate of the antibody-forming cells of the splenic foci. Eur. J. Immunol. *26*, 444–448.

Snapper, C.M., and Paul, W.E. (1987). Interferon-gamma and B cell stimulatory factor-1 reciprocally regulate Ig isotype production. Science *236*, 944–947.

Spencer, J., and Sollid, L.M. (2016). The human intestinal B-cell response. Mucosal Immunol *9*, 1113–1124.

Stavnezer, J. (1996). Immunoglobulin class switching. Current Opinion in Immunology *8*, 199–205.

Stubbington, M.J.T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A.O., Dougan, G., and Teichmann, S.A. (2016). T cell fate and clonality inference from single-cell transcriptomes. Nat Meth *13*, 329–332.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956–960.

Sundararajan, A., Sangster, M.Y., Frey, S., Atmar, R.L., Chen, W.H., Ferreira, J., Bargatze, R., Mendelman, P.M., Treanor, J.J., and Topham, D.J. (2015). Robust mucosal-homing antibody-secreting B cell responses induced by intramuscular administration of adjuvanted bivalent human norovirus-like particle vaccine. Vaccine *33*, 568–576.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat Meth *6*, 377–382.

Tas, J.M.J., Mesin, L., Pasqual, G., Targ, S., Jacobsen, J.T., Mano, Y.M., Chen, C.S., Weill, J.-C., Reynaud, C.-A., Browne, E.P., et al. (2016). Visualizing antibody affinity maturation in germinal centers. Science aad3439.

Temoltzin-Palacios, F., and Thomas, D.B. (1994). Modulation of immunodominant sites in influenza hemagglutinin compromise antigenic variation and select receptor-binding variant viruses. J Exp Med *179*, 1719–1724.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Tung, P.-Y., Blischak, J.D., Hsiao, C., Knowles, D.A., Burnett, J.E., Pritchard, J.K., and Gilad, Y. (2016). Batch effects and the effective design of single-cell gene expression studies. bioRxiv 062919.

Vandenbon, A., Dinh, V.H., Mikami, N., Kitagawa, Y., Teraguchi, S., Ohkura, N., and Sakaguchi, S. (2016). Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. Proc Natl Acad Sci U S A *113*, E2393–E2402.

Wang, B.-Z., Xu, R., Quan, F.-S., Kang, S.-M., Wang, L., and Compans, R.W. (2010). Intranasal Immunization with Influenza VLPs Incorporating Membrane-Anchored Flagellin Induces Strong Heterosubtypic Protection. PLoS ONE *5*, e13972.

Wang, G.C., Dash, P., McCullers, J.A., Doherty, P.C., and Thomas, P.G. (2012). T-cell receptor αβ diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection. Sci Transl Med *4*, 128ra42.

Wang, M., Yuan, J., Li, T., Liu, Y., Wu, J., Di, B., Chen, X., Xu, X., Lu, E., Li, K., et al. (2011). Antibody Dynamics of 2009 Influenza A (H1N1) Virus in Infected Patients and Vaccinated People in China. PLOS ONE *6*, e16809.

Wang, T.T., Maamary, J., Tan, G.S., Bournazos, S., Davis, C.W., Krammer, F., Schlesinger, S.J., Palese, P., Ahmed, R., and Ravetch, J.V. (2015). Anti-HA Glycoforms Drive B Cell Affinity Selection and Determine Influenza Vaccine Efficacy. Cell *162*, 160–169.

Wardemann, H., Yurasov, S., Schaefer, A., Young, J.W., Meffre, E., and Nussenzweig, M.C. (2003). Predominant Autoantibody Production by Early Human B Cell Precursors. Science *301*, 1374–1377.

Weigert, M.G., Cesari, I.M., Yonkovich, S.J., and Cohn, M. (1970). Variability in the Lambda Light Chain Sequences of Mouse Antibody. Nature *228*, 1045–1047.

Weinberg, A., Song, L.-Y., Walker, R., Allende, M., Fenton, T., Patterson-Bartlett, J., Nachman, S., Kemble, G., Yi, T.-T., Defechereux, P., et al. (2010). Anti-Influenza Serum and Mucosal Antibody Responses after Administration of Live Attenuated or Inactivated Influenza Vaccines to HIV-Infected Children. J Acquir Immune Defic Syndr *55*, 189–196.

Weinstein, J.A., Jiang, N., White, R.A., Fisher, D.S., and Quake, S.R. (2009). High-throughput sequencing of the zebrafish antibody repertoire. Science *324*, 807–810.

Weis, W., Brown, J.H., Cusack, S., Paulson, J.C., Skehel, J.J., and Wiley, D.C. (1988). Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. Nature *333*, 426–431.

WHO (2016). WHO | Influenza (Seasonal). http://www.who.int/mediacentre/factsheets/fs211/en/

Wilson, P.C., and Andrews, S.F. (2012). Tools to therapeutically harness the human antibody response. Nat Rev Immunol *12*, 709–719.

Wilson, I.A., Skehel, J.J., and Wiley, D.C. (1981). Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. Nature *289*, 366–373.

Winkelhake, J.L. (1978). Immunoglobulin structure and effector functions. Immunochemistry *15*, 695–714.

Woof, J.M., and Mestecky, J. (2005). Mucosal immunoglobulins. Immunological Reviews *206*, 64–82.

Wrammert, J., Smith, K., Miller, J., Langley, W.A., Kokko, K., Larsen, C., Zheng, N.-Y., Mays, I., Garman, L., Helms, C., et al. (2008). Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. Nature *453*, 667–671.

Wrammert, J., Koutsonanos, D., Li, G.-M., Edupuganti, S., Sui, J., Morrissey, M., McCausland, M., Skountzou, I., Hornig, M., Lipkin, W.I., et al. (2011). Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. J Exp Med *208*, 181–193.

Wrammert, J., Onlamoon, N., Akondy, R.S., Perng, G.C., Polsrila, K., Chandele, A., Kwissa, M., Pulendran, B., Wilson, P.C., Wittawatmongkol, O., et al. (2012). Rapid and massive virus-specific plasmablast responses during acute dengue virus infection in humans. J. Virol. *86*, 2911–2918.

Xu, Z., Zan, H., Pone, E.J., Mai, T., and Casali, P. (2012). Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. Nature Reviews. Immunology *12*, 517–531.

Yamane-Ohnuki, N., Kinoshita, S., Inoue-Urakubo, M., Kusunoki, M., Iida, S., Nakano, R., Wakitani, M., Niwa, R., Sakurada, M., Uchida, K., et al. (2004). Establishment of FUT8 knockout Chinese hamster ovary cells: an ideal host cell line for producing completely defucosylated antibodies with enhanced antibody-dependent cellular cytotoxicity. Biotechnol. Bioeng. *87*, 614–622.

Young, J. (2002). Development of a potent respiratory syncytial virus-specific monoclonal antibody for the prevention of serious lower respiratory tract disease in infants. Respiratory Medicine *96*, S31–S35.

Yu, X., Tsibane, T., McGraw, P.A., House, F.S., Keefer, C.J., Hicar, M.D., Tumpey, T.M., Pappas, C., Perrone, L.A., Martinez, O., et al. (2008). Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. Nature *455*, 532–536.

Zhou, F., Li, X., Wang, W., Zhu, P., Zhou, J., He, W., Ding, M., Xiong, F., Zheng, X., Li, Z., et al. (2016). Tracing haematopoietic stem cell formation at single-cell resolution. Nature *533*, 487–492.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., G&#xf3, Rczak, K., Klamecka, K., Figlerowicz, M., and Siatkowski, I. (2015). The Impact of Normalization Methods on RNA-Seq Data Analysis. BioMed Research International *2015*, e621690.