(THIS PAGE LEFT INTENTIONALLY BLANK)

Next-generation *ABO* Genetics and Genomics

Patrick Keolu Ozer Fox

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Debbie Nickerson, Chair

Jill Johnsen

Evan Eichler

Program Authorized to Offer Degree:
Genome Sciences

(THIS PAGE LEFT INTENTIONALLY BLANK)

University of Washington

Abstract

Next-generation *ABO* genetics and genomics

Patrick Keolu Ozer Fox

Chair of the Supervisory Committee:
Dr. Debbie Nickerson
Department of Genome Sciences

Accurately cross matching units of blood based on blood type is essential for successful transfusion therapy. ABO is the most clinically relevant blood group in transfusion therapy due to the presence of naturally occurring ABO antibodies. Failure to correctly match ABO blood type can cause fatal transfusion reactions even in transfusion naïve individuals.

The *ABO* gene commonly encodes two different forms of a glycosyltransferase which adds A or B sugars (N-acetylgalactosamine for A or α-D-galactose for B) to the H-antigen substrate. Single nucleotide variants (SNVs) and insertion-deletions (indels) in the *ABO* gene affect function at the molecular level by altering the specificity and efficiency of the enzyme for specific sugars (leading to the A1, A2, and B blood types) or by knocking out gene function to generate the O blood type. Thus, variation in A, B, or O serological phenotype is the result of genetic variation in the coding portion of the *ABO* gene. Currently, approaches to genotype *ABO* are limited because *ABO* is a complex locus with a large number of functional haplotypes that can lead to the A, B, or O phenotype.

In addition to many other factors, ABO blood type plays a role in determining an individual's risk for multiple common complex diseases including the number one and number two causes of death in the United States: cardiovascular disease and cancer. However, the influence of specific *ABO* types and *ABO* subtype variants, such as the A1 and A2 haplotype/subtypes, on common complex disease risk has not yet been fully explored.

In this dissertation, I directly explore the many forms of variation in the *ABO* gene in diverse human populations using multiple next-generation human genome sequencing datasets, while simultaneously addressing the limitations of both traditional serological methods and existing genotyping methods designed to determine ABO blood type from variation found in the *ABO* gene. I then discuss strategies and limitations of developing an automated approach to call high resolution phased ABO blood types from NGS data.

The methods and analyses outlined in this dissertation can be used to generate higher resolution blood type and subtype calls leveraging the variation and phenotypes within large scale NGS populations based to explore the relationships between rare and common *ABO* variants, *ABO* haplotypes, and subtypes with common complex disease-related phenotypes (i.e., cardiovascular disease, cancer, and type 2 diabetes). My hope is that the NGS tools developed in this thesis will be used to create a more comprehensive understanding of common complex disease etiology in the future.

FOREWORD

Over the last five to six years, it has been an honor to be on the ground floor for multiple paradigm shifts in the Nickerson lab: First the NHLBI exome sequencing project (2010), followed by the Center for Mendelian Genomics (2013), then the BloodSeq project (2014), and most recently the completion of thousands of whole human genomes through the Trans-Omics for Precision Medicine (TOPMed) Program (2016). My tenure in graduate school has been challenging, humbling, and insightful on many levels. We (the genomics/biomedical community) are on the brink of a revolution in medicine. For the first time in medical history we are approaching a monumental singularity in medicine that for thousands of years has been inherently reactive; we finally have the biological knowledge, computational potential, and emerging technology to begin to predict and prevent both rare and common chronic complex diseases using genome-sequencing technologies.

President Barack Obama has spoken supportively about this new direction in "personalized" and "precision medicine." Federal policy makers have allocated a budget of $215,000,000.00 to establish the "Precision Medicine Initiative (PMI)," an effort to sequence one million whole human genomes over the next two years (projected to be completed by 2018). In their words, the PMI represents an, "emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. While some advances in precision medicine have been made, the practice is not currently in use for most diseases." The future of this type of medicine is looking bright. Genomics, once thought of as an emerging discipline, has grown to promise immense potential. Yet I still have some very real concerns about the broad application of Precision Medicine in the future.

I hope to prove to my committee members that in order for the genomics community to broadly apply Precision Medicine it will need to rely on, (1) basic science, in the form

of proof-of-concept contributions that utilize emerging technology (i.e., next-generation sequencing) to revisit classic biological problems, long thought to be solved (e.g., *ABO* blood typing) and (2) diversity, which matters more than ever as our country becomes more multicultural. If we (in the U.S.) are going to sequence 1,000,000 American genomes, our current efforts in obtaining equal representation for under-represented minority populations in large-scale population-based screenings need to reflect this diversity. We as a community are not doing enough to discover human genetic variation in all of its forms (SNVs and CNVs, for example) and especially in underrepresented populations that might otherwise be left out of the predictive and preventative medicine revolution.

My hope is that this thesis, "*Next-generation* ABO *Genetics and Genomics*," will be a small contribution to the existing literature in genomic technologies and health-disparities research. In order to democratize Precision Medicine we need to understand human genetic variation in all populations, including those that are in the minority. I believe the best place to begin is with one of the most well-characterized genes in the human genome, *ABO*. Complementing serological ABO-based blood typing with next-generation sequence-based ABO blood typing is but one part of the effort to realize the true potential of Precision Medicine. Let the next-generation of *ABO* genetics and genomics begin.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTAL FIGURES

Figure Number Page

LIST OF TABLES

LIST OF SUPPLEMENTAL TABLES

ACKNOWLEDGMENTS

Words cannot explain (though I will try) how grateful I am to the many individuals who have offered academic, financial, and emotional support during the preparation of this dissertation. I first thank my Ph.D. supervisor, Debbie Nickerson: you have always challenged and brought out the best in me. You exposed me to a new world of science, ideas I never knew were possible, and surrounded me with exceptional scientists and experiences that have helped me grow into the scientist I am today. Your dedication to increasing diversity in genome sequencing studies has inspired me. It was an honor to write and win an official NIH grant with you. We received a perfect score, probably the first and last in my career — and I published my first, first-author paper with you as the corresponding author. You are a force; and one of the most generous people I have ever met. I will strive my whole career to emulate your success, attention to detail, and focus.

The remaining members of my committee each played different but important roles in my training. I am grateful to Evan Eichler for allowing me to rotate in his laboratory. As a first-year student, I walked into your office with some wild ideas. Rather than turning me away, you helped me re-direct my purpose, imagination, and focus on bite-sized projects that eventually grew into searching for SV in *ABO,* my first paper. I found my first taste of success in your laboratory working with Santhosh Girirajan. It is an honor to be included in the author list with you on multiple papers. Jill Johnsen opened her arms and allowed me into the ABO world. Your warmth as a scientist and medical doctor is something I want to emulate throughout my career. You steered me away from some questionable decisions and indulged me in my deep dives into the hematology world which I knew nothing about. It has been an honor to be included on multiple publications with you over the last five years. Alex Reiners's positive sprit and tactful constructive criticism of my work has improved the quality of each project we have

worked on together. Mike Bamshad allowed me to rotate in his laboratory my first quarter in graduate school. Your work on CCR5 and balancing selection was a significant factor in why I choose to come to UW in the first place. Finally, Phil Green never once turned me away when I knocked on his office door. A man of few words — when you speak, the world listens. Stan Fields, though you were not on my committee, you have been a wonderful advisor to me over the past five years. You have influenced my character both as a scientist and a man.

While I have benefited enormously from interactions and collaborations with every member of the Nickerson lab, I would like to single out several individuals for contributions that merit particular thanks. Adam Gordon, although you are a year younger than me, I looked up to you in many ways as the senior grad student in the Nickerson lab. You are the Wikipedia for genomics knowledge. I am grateful for your advice throughout my tenure in graduate school. You always made time to sit down and triple-check my slides for presentation. I am grateful for your time and thoughtfulness. Ian Stanaway took me on as an inexperienced rotation student and was chiefly responsible for exposing me to many ideas concerning computed science. His attention to detail, his focus on algorithm development, and his willingness to share his exceptional skills, both in the lab and on the BBQ grill, is inspirational. David Crosslin taught me to never to judge a book by its cover. You are pragmatic, and you have a rare gift for exposing the elegance and simplicity of hardcore mathematics. You showed me the world of statistics and disease modeling. I am still trying to process ideas from our spirited conversations on regression analyses. Tristian Shaffer, you always found ways to challenge my imagination. You are a wealth of knowledge from dirt bike engine assembly, to genome assembly. You have a gift for distilling complicated ideas in bioinformatics to their most simple components. Jason Underwood, you might have best hands I have ever seen on the bench. Your creativity there inspired me to try my own “hands” even after months of frustrating failure you encouraged me to try, and then try again. When that didn’t work, you went back to the drawing board with me, to try out a new

perspective. Marsha Wheeler, a wonderful addition to the Nickerson laboratory and my partner-in-crime while working on the Blood-seq project. Your contagious smile always managed to lift my spirits when I was having a bad day, *muchas gracias*. And finally, Colleen Davis. None of this would be possible without you. Even with two young daughters and a full plate of projects you always made time to copy-edit my F31, papers, and letters of recommendation — you are greatly appreciated.

I am also grateful to my good friends and collaborators in the Eichler lab. A special thanks goes out to Santhosh Girirajan. You made me believe in myself. You are easily one of the most hard-working people I have ever met. I am over the moon happy for you and your new family. Brad Coe, you are a wealth of knowledge. Easily one of the friendliest people I know, inside or outside of genomics. I appreciate our conversations over both coffee and beer. Thanks also to Tonia Brown, Megan Dennis, and Micheal Duyzund.

I am also grateful to my neighbors/friends in the Shendure lab. Thanks to Jay, Martin Kircher, Jerrod Schwartz, Akash Kumar, Aaron Mckenna, Greg Findlay, Andrew Adey, Joe Hiatt, and Matthew Snyder. Working next to a group of some of the most dedicated people in genomics is both motivating and inspiring. Thank you for entertaining many of my conversations.

I have been very lucky to complete my doctoral research in the Department of Genome Sciences at this moment in time, surrounded by an amazing group of faculty, administrators, and trainees — the very reason I was attracted to the department in the first place. Thanks to my “day-one homie,” Billy Edelman. You are one of the reasons I chose to come to genome sciences. You are a creative force and in many ways my muse. I look forward to the next “wrinkle” in your career. Thanks to my classmates: Josh Burton, Xander Nuttle, and Matt Rich. A huge thank you to Brian Giebel. The day you called me to come an interview at genome sciences was one of the best moments of my life. Our conversations have filled me with confidence. Thanks, to Dawn Counts, Lisa Boader, and Maureen in the grants office.

I thank my Ohana for believing in me; the reason I am here is both because of, and for YOU. To my mother — I learned to work hard through your example. I am proud to be Hawaiian because of you. When I was content with my high school diploma, you told me that would never be enough. To my father: you inspired me to be a scientist. As a child, I remember asking you why the sun was following us. Your response was priceless, "think about it." Your infectious curiosity has inspired me to "think" about the mysteries of the world for a living. To my Aunty Aloha: your work as a nurse at the native health medical center influenced and inspired me to focus on the injustices taking place in the public health world. To my cousins, Makanai, Naupualani, and Kukaiau — Mahalo nui loa for never letting me take myself too seriously. To the Taitingfong clan: many mahalos for accepting me into your family and allowing me to spend holidays at your home. To my god-parents; Sandy and Harold. It is in large part through your honesty, and countless dinner conversations concerning politics, activism, and social justice issues that my moral values were forged. Thank you for helping me with this very document. And to my grandparents, both of whom passed away while I was in graduate school: I know you would have loved to share this moment – this achievement is for you. I thank the love of my life, Riley Taitingfong. My rock. Before I met you I was seriously considering dropping out of this Ph.D. program. You inspired me to wake up earlier each day and work hard to pursue my dreams. You most of all made me believe in myself.

Finally, to those individuals who graciously donated your genomes to science. Without you none of this work would be possible.

DEDICATION

This work is dedicated to my mother:

Even as a single mom, you made time for late nights working on science projects in elementary school. We brought home that blue ribbon once or twice…

and

my ʻOhana:

Ohana mua. I aloha ia oukou a pau

(THIS PAGE LEFT INTENTIONALLY BLANK)

Chapter 1

INTRODUCTION

The *ABO* blood group system (Landsteiner, 1900) is one of the oldest known genetic loci in humans. Prior to the discovery of the ABO blood group, early attempts at transfusion therapy resulted in grave consequences. Characterizing variation in the *ABO* gene is important in transfusion and transplantation medicine because variants in *ABO* have significant consequences with regard to recipient compatibility. While many believe that the genetics of the *ABO* locus is well understood, little is known about the impact of novel, rare variation in the *ABO* gene. In this chapter, I briefly review the history of the ABO blood group, genetics, biochemistry, ABO blood typing, and *ABO*'s importance in transfusion medicine/ disease impact. I also detail the types of information that ABO blood typing can provide to clinicians and patients, and discuss the potential directions for and implications of further investigation into genetics and genomics related to this medically important locus.

1.1 A brief history of the ABO blood group

Karl Landsteiner discovered the ABO blood group system in 1900 (Landsteiner, 1900). Landsteiner's initial experiments separated the cell components and the fluid (or sera) of blood from different individuals (including his own blood) and he then mixed these together in various combinations. His observation was that some combinations of cells and sera led to clumped or agglutinated red blood cells (RBCs), while others did not. He observed three agglutination patterns that he designated as A, B, and C. He later changed the name of blood group "C" to blood group "O." Less than a year later, using the same logic, Decastello and Sturli would describe the "AB" blood type (Decastello and Sturli, 1902).

Landsteiner theorized RBCs contained two different "markers," or "indicators," that were capable of reacting with naturally occurring elements found in blood serum, i.e.,

anti-A and anti-B agglutinins. Thus, individuals with the A type of RBCs would have naturally occurring anti-B agglutinins capable of agglutinating or clumping B, and AB "type" RBCs but not their own A RBCs (*vice versa* for blood type B.) In the case of O RBCs, both anti-A agglutinins and anti-B agglutinins are naturally present making it capable of agglutinating A, B, and AB RBCs. Finally, individuals with AB-type RBCs have neither anti-A nor anti-B agglutinins present. These findings and conclusions would later be called Landsteiner's law (see Figure 1.1).

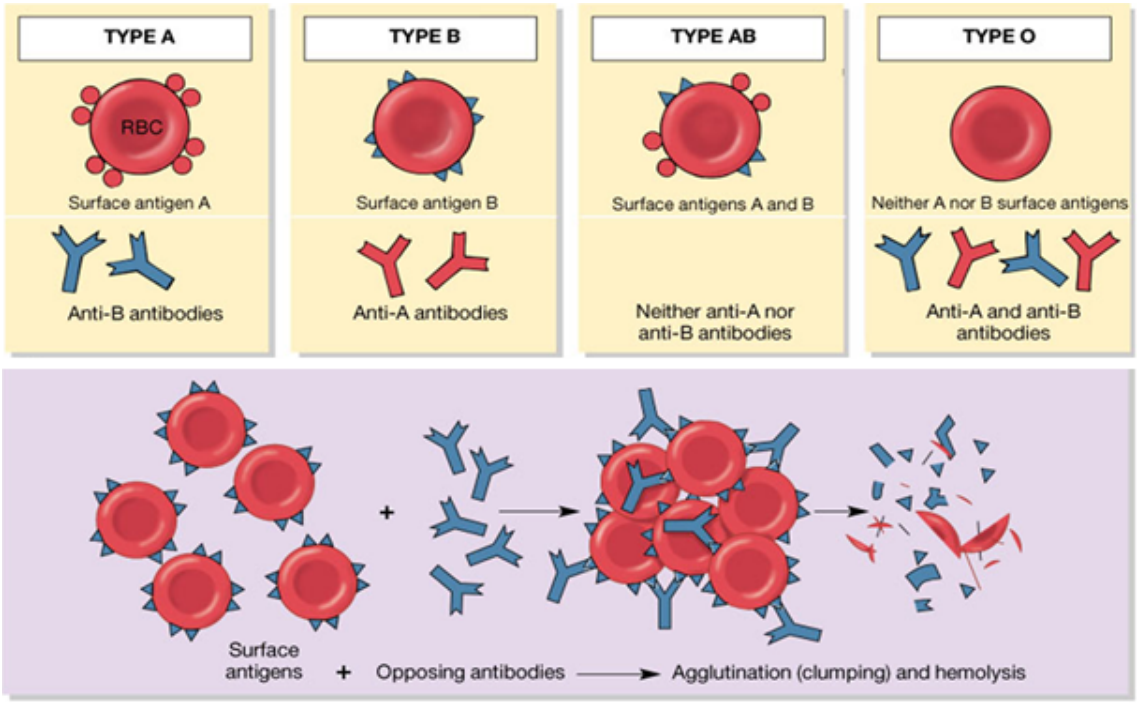

Figure 1.1: **Landsteiner's law.** (1) If an "agglutinogen," later called an antigen, is present on the red blood cell, the corresponding "agglutinin", or antibody, must be absent in the sera. (2) If an antigen is absent on the red blood cell, then the corresponding antibody must be present in the sera. (Modified from Olsson and Storry, 2009.)

Concurrent with Landsteiner's work, Ottenberg and Epstein (1908) were the first to suggest that ABO blood groups might be inherited (Ottenberg and Epstein, 1908). Their finding was confirmed two years later by Von Dungren and Hirszfeld (1910), who used

kinship investigation of 72 families with 102 children (i.e., pedigree relationships), to demonstrate that the inheritance of ABO obeyed Mendel's laws, and thus its utility as both a forensic and paternity test (von Dungren and Hirszfeld, 1910). The two researchers hypothesized that the A and B antigens were produced by two independent dominant alleles. However it wasn't until 1924 that Felix Bernstein (1924) hypothesized that there were actually multiple alleles at one locus; this would later be called the "one gene three locus model" (Berstein, 1924). By applying Hardy-Weinberg principles, Berstein also hypothesized that the A and B alleles were co-dominant against the recessive O allele (Crow, 1993).

In 1926 Putkonen and Lehrs demonstrated that A and B antigens were not only found on the surfaces of RBCs but were also in water-soluble form via secretions such as semen and saliva. Moreover, they discovered that the ability to secrete these antigens was genetically independent from the *ABO* gene but also inherited via the classical mode of Mendelian inheritance (Putkonen, 1930). In 1943 Walter Morgan and his Ph.D. student H.K. King pioneered techniques for isolating and characterizing a component of hog gastric mucin with considerable blood group A activity measured by isoagglutination inhibition and erythrocyte lysis experiments (Morgan W. T. and King, H. K. 1943). Until then, no one understood what the A antigen was. Clearly, it was found on the cell surface or in soluble form, and involved carbohydrates, lipids, and amino acids (Mulloy, B. et al., 2014). The new "A substance" could degrade under heat or alkaline and then lose its ability to isoagglutinate; at the same time, it had a somewhat enhanced ability to inhibit cell lysis (Mulloy, B. et al., 2014).

In the 1960's the carbohydrate structures that constitute the ABO(H) antigen were confirmed independently by Morgan at the Lister Institute in England and Elvin Kabat and Winifred Watkins at the Columbia University; they also determined the chemical nature of the A, B, and H antigen substances or "markers" first described by Landsteiner (Watkins, Kabat, and Morgan, 1966); they used alkaline to cleave the A, B, and H sugar

chain fragments from various tissue types followed by purification and characterization of their differences via paper chromatography, i.e., separation by polarity (Watkins, Kabat, and Morgan, 1966). One key result Kabat highlighted was that H antigens were found in high abundance in individuals with the O blood type. However, the researchers' most important contribution was the discovery that the ABO blood-group gene antigen products were not the primary gene products at all but in fact were the product of enzymatic reactions building the carbohydrate chains (Watkins, Kabat, and Morgan, 1966).

In 1976 the metabolic pathway responsible for the biosynthesis of the A, B, and H antigens was established (reviewed by Watkins, 1981) and the *ABO* gene was localized and assigned to the distal end of the long arm of chromosome 9 (9q34) using a precise type of linkage group mapping called deletion mapping (Ferguson-Smith et al., 1976). With the development of forensic DNA analysis technology in the late 1980's, the possibility of "DNA fingerprinting" became available to characterize the A, B, and O alleles (Jeffreys et al., 1985). In 1990, Clausen et al. purified the soluble form of the A antigen transferase.

Fumiichiro Yamamoto at the University of Washington then built upon the glycolsyl-transferase purification experiment (Clausen et al., 1990), by using reverse transcriptase (Baltimore et al., 1981) to create complementary DNA libraries (cDNA); he adopted the restriction enzyme-based assessment of single locus polymorphisms to isolate and describe the sequence of the first cDNA clone of the A glycosyltransferase (Yamamoto et al., 1990.) Yamamoto followed up with the B glycosealtranferase and O protein, to conclusively demonstrate the molecular basis and synthesis of the A and B antigens (Yamamoto et al., 1990).

Starting in the late 1990's-2000's, major developments occurred in assessing the structure/function relationship of ABO glycosyltransferases (Yamamoto et al., 1990). A and B transferases were cloned and shown to differ in four critical amino acid residues

(Arg/Gly 176, Gly/Ser 235, Leu/Met 266 and Gly/Ala 268, respectively) out of their total of 354 amino acid residues (Patenaude et al., 2001). The combination of the identification of amino acid substitutions between A and B transferases and the growing number of genetic variants and their corresponding amino acid substitutions identified in the *ABO* gene after the initial cloning experiments by Yamamoto et. al. (1990) have led to a more comprehensive understanding of the genotype-phenotype relationship *via* the effect of *ABO* variation on glycosyltransferase structure/function. Yamamoto isolated genomic DNA clones encompassing ~30 kb surrounding the *ABO* locus; the locations of the exons were then mapped and the nucleotide sequences of the exon/intron boundaries determined (Yamamoto et al., 1995).

In the early 2000's Evans et al. (2001) used crystallography, i.e., the determination of the 3D-structure of the glycosyltransferase *via* molecular conformation of biological macromolecules to create a comprehensive understanding of the structure/function relationship of the ABO glycosyltransferases for both the A and B glycosyltransferase domain structures, highlighting active-site domains, including acceptor and donor sites on the ABO glycosyltransferase (Patenaude et al., 2001). Once both the A and B alleles were identified and those essential amino acid substitutions were linked to alterations in the structure/function of critical domain structures (e.g., the active site domain), the next step forward would focus on understanding the effects of both common and rare variation in the *ABO* gene on the structure/function of the glycosyltransferase.

1.2 The *ABO* gene

The human *ABO* locus, spanning ~18 kb, lies on chromosome 9q34 and encompasses seven coding exons composed of a 1062-bp open-reading frame. The exons range in size from 28 to 688 bp, with the majority of the coding sequence lying in exon 6 and 7. Exons 1 through 5 encode the amino terminal part, a transmembrane region, and 9 percent of

the catalytic domain. Exons 6 and 7 encode 77% of the protein and 91% of the catalytic active part (Olsson, 2009), while the introns range in size from almost 13 kb to 554 bp.

For a detailed illustration of the organization of the *ABO* gene, its scale, and sequence comparisons of various alleles in exons 6 and 7, see Figure 1.2.

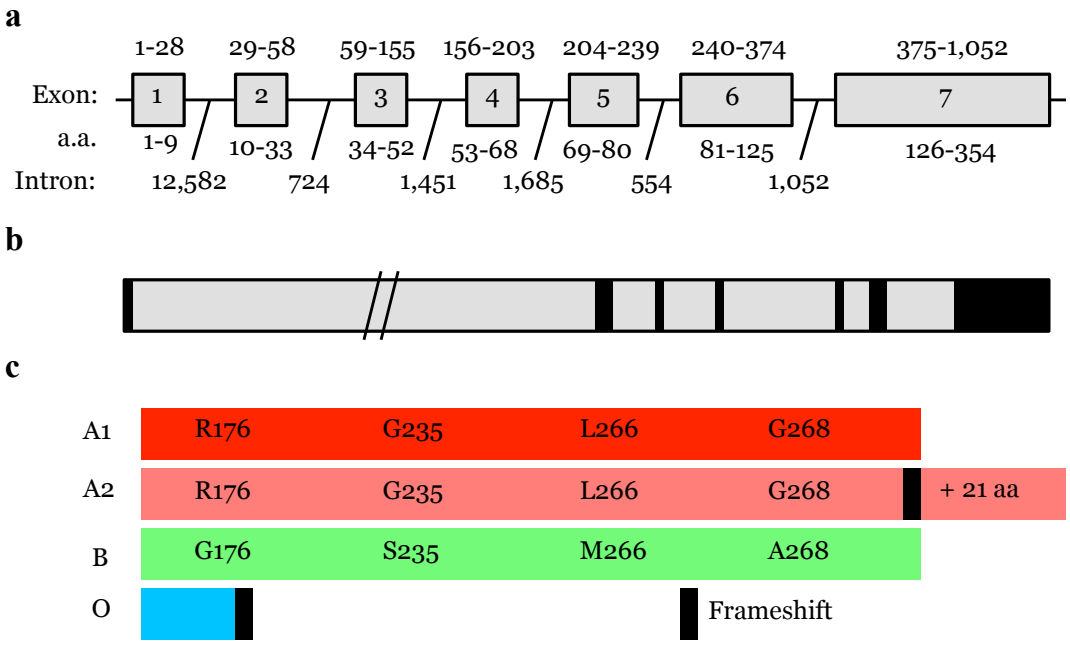

Figure 1.2: **Organization of the *ABO* gene.** (a) *ABO* is a negative-strand gene. Here it is represented in its linear form (from left to right). The seven exons and six introns are not drawn to scale. The numerals above the boxes represent the first and last nucleotides (bp) of the coding region in each exon, and those below the boxes show the corresponding amino acid (a.a.) numbers. The size of each intron is indicated with a thin oblique bar (bp). (b) The *ABO* gene is drawn to scale (except intron 1); exons are black and introns gray. (c) The A and B alleles encode an a3-N-acetylgalacto-saminyl-transferase (A transferase) and an a3-galactosyltransferase (B transferase) respectively. The A enzyme transfers a-GalNAc from UDP-GalNAc to the H antigen. The B enzyme transfers a-Gal from UDP-Gal to the H antigen. Sequence comparison at the *ABO* locus in A, B, and 0 subjects has revealed 4 amino acid differences between the A and the common B allele (R to G, G to S, L to M and G to A). These differences account for the different substrate specificity of the A and B glycosyltransferases. The 0 allele has neither transferase activity because of a loss of function (LOF) frameshift in exon 6. The A2 allele is the result of a read-through frameshift at the end of exon 7 and the addition of +21 amino acids, which alters the efficiency specificity of the A2 transferase. (Adapted from Storry and Olsson, 2009)

The *ABO* gene encodes a glycosyltransferase composed of 354 amino acids, which adds different sugars (N-acetylgalactosamine for A and α-D-galactose for B) to the H antigen substrate (Evans et al., 2001). Single nucleotide variants (SNVs) in the *ABO* gene affect the function of this glycosyltransferase at the molecular level by altering the specificity and efficiency of this enzyme for these specific sugars (Yamamoto et al., 1990). The products result in A or B blood-group-specific antigens. O is caused by loss of function (LOF) variants in the *ABO* gene. A common cause of O is an *ABO* exon 6 deletion which induces a frameshift and creates a premature stop codon (nucleotides 352-354), resulting in a truncated (117 amino acids) protein deprived of any glycosyltransferase activity (Evans et al., 2001). While the majority of O haplotypes result from the exon 6 deletion, there are less common LOF O haplotypes (called non-deletional O haplotypes) that do not include LOF nonsense variation in exon 6 (Yazer et al., 2008).

Normally, individuals have two copies of the *ABO* gene, one gene inherited from each parent. Differentiating the variants on one copy of the *ABO* gene from the other is known as haplotyping (Clark et al., 1990). The ABO blood type is considered a classic example of a co-dominant trait inheritance in the human genome (see Figure 1.3 for a depiction of co-dominant inheritance). Relating *ABO* genotype to blood group antigen phenotype requires the analysis of individual variants and the haplotype inherited from each parent. Each haplotype is composed of both common and rare alleles in the *ABO* gene. In Chapter 2, I will further detail the importance of haplotyping in the analysis of the ABO locus.

Because ABO is a co-dominant phenotype when both the A and B alleles are inherited (one from each parent), it results in a less-common phenotype, the AB blood type. A and B alleles are both dominant over the recessive null (recessive) O allele. Thus, individuals who are homozygous for haplotype “A” (A/A) as well as individuals who are heterozygous for haplotype “A” and “O” (A/O) both type as blood type “A” using

traditional serological methods (see blood typing in section 1.(3). The majority of these haplotype motifs are composed of multiple missense variants occurring simultaneously on the same haplotype block resulting in critical amino acid substitutions located in the active site domain of either the A or B glycosyltransferase (Evans et al., 2001). Specifically there are six critical locations located in both exons 6 and 7 that define most common A, B, and O haplotypes (see Figure 1.2.c).

However, the effect of non-synonymous variation in *ABO* should be viewed as a phenotypic gradient of variation that is manifested as differences in glycosyltransferase activity, such as weak agglutination phenotypes, e.g., the A2 subtype (See section 1.(3). In the case of the A2 phenotype, the A2 glycosyltransferase doesn't convert the H antigen to the A antigen with high efficiency because of a rare frameshift at the terminal end of exon 7, resulting in a read-through, and the addition of 21 amino acids, thus creating a more cumbersome glycosyltransferase (Yamamoto et al., 1992). The result is fewer A2 antigen sites on the surface of RBCs and more exposed H antigen sites when compared to the higher number of A antigen sites and, conversely, less exposed H antigen sites on the traditional A1 RBCs (see Figure 1.5).

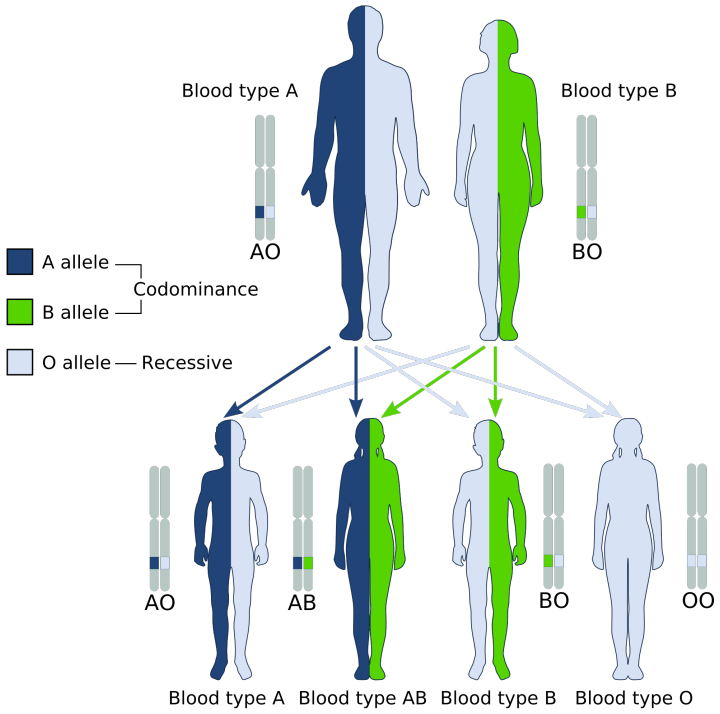

Figure 1.3: **A and B are codominant, giving the AB phenotype.** ABO blood groups are inherited from both parents. The ABO blood type is controlled by a single gene (the *ABO* gene) with one copy inherited from each parent, two haplotypes total. The *ABO* gene encodes a glycosyltransferase enzyme that creates carbohydrate structures on red blood cells and other tissues which is highly antigenic. (Modified from Klug et al., 1997.)

1.3 ABO biochemistry and structure

A and B antigens share the same structure with the exception of a terminal sugar bound by a α1-3 glycosidic linkage to galactose. (Yamamoto, 2014). For the A antigen, the terminal sugar is a N-acetylgalactosamine (Gal-NAc) and for the B antigen the sugar is α-D-galactose (Gal). If either the A or B terminal sugars are eliminated from the common structure, their corresponding antibodies (anti-A and anti-B) lose their reactivity (Yamamoto, 2014). The natural precursor antigen for the attachment of the terminal Gal/Gal-NAc sugar is called the H antigen. The H antigen is defined by an alpha-1,2-fucose residue synthesized by fucosyltransferases encoded by either the the *FUT1* or

FUT2 genes. For an illustration of the sequence of the generation of ABO antigens, see Figure 1.2. The H antigen is essential for the A and B glycosyltransferase to recognize it as the acceptor and transfer either the Gal or Gal-NAc to the terminal Gal (Yamamoto, 2014). Various *FUT1* and *FUT2* alleles can produce modified or H-deficient phenotypes as a result of both missense and LOF frame shifts (Yamamoto et al., 2014).

The Bombay phenotype is the total absence of the H antigen (the products of the *FUT1* and *FUT2* genes) on RBCs and secretions, independent of ABO blood type (Yamamoto et al., 2014). H antigen deficiency, i.e., the "Bombay phenotype," is found in one of 10,000 individuals in India and one in a million people in Europe. There is no ill effect with being H deficient; however, if a blood transfusion is needed, people with this blood type can receive blood only from other donors who are also H deficient. A transfusion of group O blood can trigger a severe transfusion reaction due to anti-H antibodies. If the H antigen, precursor of the ABO blood group antigens, is not produced, the ABO blood group antigens are also not produced.

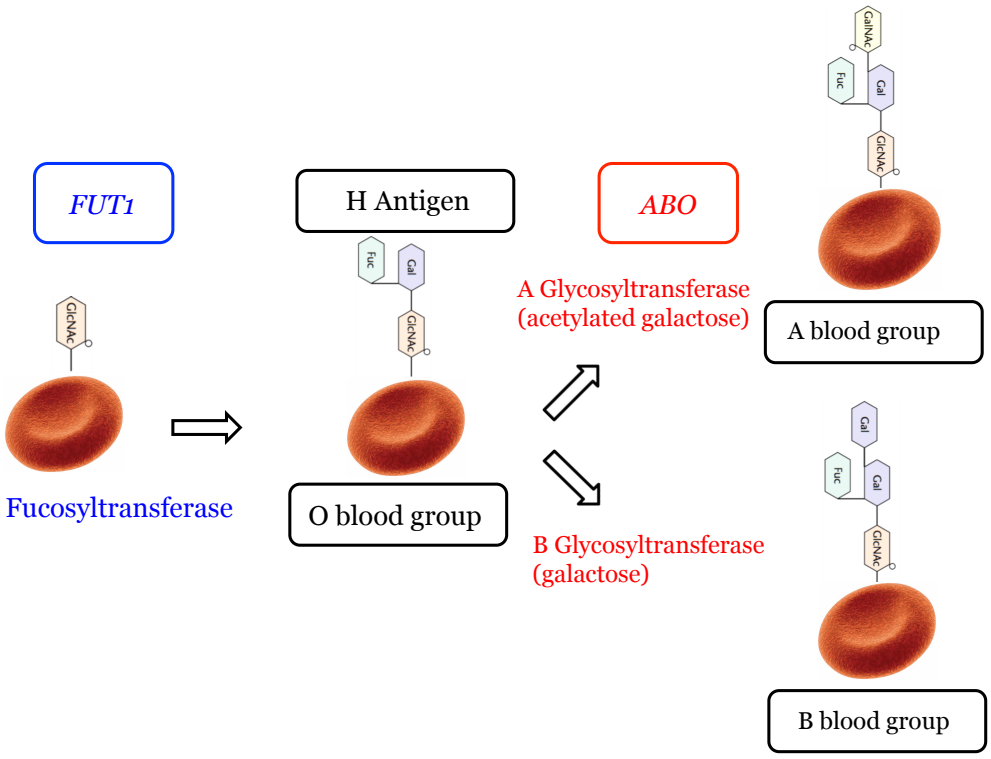

Figure 1.4: **A, B, and H antigen biosynthesis.** The synthesis of A, B, and H antigen synthesis occurs in a two-step process. (1) The *FUT1* and *FUT2* genes are responsible for synthesizing a fucosyltransferase enzyme that attaches the H antigen to glycoproteins on the surface of RBCs (and other tissues). (2) Depending on the variation in ABO, the ABO gene synthesizes a glycosyltransferase that attaches either an A or B sugar to the H antigen, thus creating the A or B antigen on the surface of the RBC.

The antigen structures detailed above decorate carbohydrate structures with variable length. Depending on the disaccharide precursor core chain, where A, B, and H determinates are synthesized, they can be divided into six different types (Yamamoto et al., 2012). The internal reducing ends of these precursors are bound to many different types of carrier molecules (generally denoted by the letter "R"). These include oligosaccharides, glycolipids, or glycoproteins (Clausen and Hakomori, 1989). Generally, types 1-4 are found on RBCs with Type 2 being the most common on those cells. Type 6 are present as free oligosaccharides and on some tissues including renal veins and intestinal cells (Bjork et al., 1987; Holgersson et al., 1990). Type 5 is synthetically derived

and was notably utilized in the categorization of monoclonal antibodies (Oriel et al., 1990). These antigen structures can be found on cell membranes bound to embedded glycoproteins and glycolipids. They can also be part of glycoconjugates suspended in fluids as plasma and secretions. Finally, they can be found as free oligosaccharides without any protein or lipid carrier (Yamamoto, 2012).

Although A, B, and H antigens were discovered on RBCs, they are also present on many other tissues. As a result, A, B, and H antigens are also called histo-blood group antigens. In whole blood, platelets and leukocytes also present A, B, and H antigens in variable amounts, depending on the blood group and cell type. A, B, and H antigens have been detected on endothelial cells, epithelial tissue, including lung and gastrointestinal tract tissues, and the lining of urinary and gestational tracts. The presence and quantity of these antigens (A, B, and H) on various tissues is relevant for blood cell (whole blood and blood component, e.g., red blood cells and platelets), and organ transplant success (reviewed in [Ravn and Dabelsteen, 2000]). For example, A2 organs have lower cell surface expression of the A antigen, and transplantation of living donor A2 renal allografts into non-A recipients can yield excellent long-term allograft survival, which expands the potential living donor pool for non-blood group A recipients (Sorensen et al., 2001).

1.4 *ABO* subgroups and weak hemagglutination

Multiple weak hemagglutination subtypes have been characterized for the A and B blood groups (Yamamoto, 2014). These are called a weak hemagglutination phenotype and account for a smaller percentage (20% of total A, in individuals of European ancestry [Yamamoto et al., 1992]) of both the A and B blood types discovered. The first recorded weak hemagglutination subtype was the A2 subtype described by von Dungren et al. (1910) after they conducted experiments characterizing differences in the amount of A antigen expression observed in multiple-A blood type individuals (von Dungren and

Hirszfeld, 1910). A1 is a subtype of the A blood group and has a distinct clumping pattern than A2 (Rochant et al., 1976): nearly 80% of individuals of European ancestry with the A blood type have the A1 subtype (Yamamoto et al., 2014). The structure and function of A2 glycosyltransferases are significantly different from A1 glycosyltransferases (Yamamoto, 2014) and the densities of antigens on A1 RBCs are four-fold greater than that of A2 RBCs (see Figure 1.5). At the genetic level, this difference is the result of a single base deletion near the 3' terminus of the *ABO* gene, thus resulting in a read-through frameshift at the end of exon 7 and the addition of +21 amino acids, which alters the efficiency and specificity of the A2 transferase (Yamamoto et al., 1992).

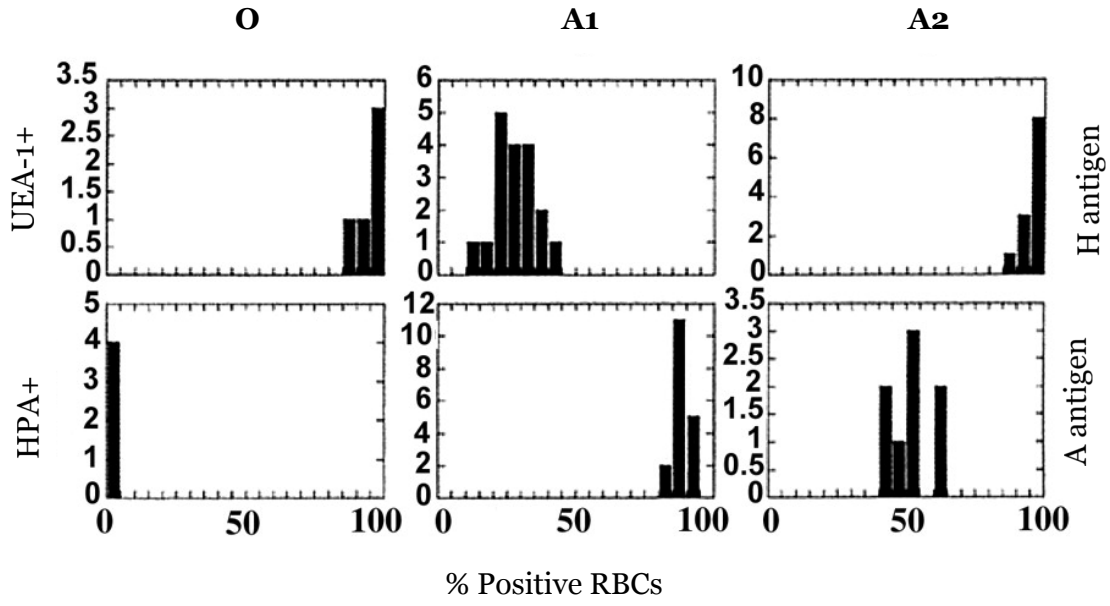

Figure 1.5: **Comparing antigen density of A1 and A2 RBC surfaces.** H and A antigen expression (FITC-UEA1) on group O, A1, and A2 donor RBCs by flow cytometry. The distribution of the H antigen, H, panel 1 (% UEA1+) and A antigen, A, panel 2 (% HPA+) on RBCs in a population of group O, A1, and A2 donors. Note that H antigen is increased on A2 RBCs (far right). (Adapted from Cooling et al., 2005.)

A2 is not the only A subtype: type A subtypes that exhibit weak expression of the A antigen include A2, A3, Aend, Afinn, Abantu, Ax, Am, Ay, and Ael (Yamamoto, 2012).

The B blood type is found at a lower frequency among multiple human populations, and fewer weak B phenotypes have been identified (Yamamoto et al., 1992). However they have been classified similarly to the A blood group subtypes (B3, Bx, Bm, Bel, and Bw). While traditional serological methods are capable of identifying some ABO subtypes, it is difficult to consistently identify weak hemagglutination phenotypes (Yamamoto et al., 1992). One complication is that A2 and A2B individuals may produce anti-A1 antibodies, which can cause apparent discrepancies when performing blood type testing. Consequently, when assigning and confirming subgroups, it is routine practice to request further tests — e.g., more in-depth characterization than forward or reverse typing such as Dolichos biflorus agglutinin (DBA) lectin testing or genetic analysis of the *ABO* gene (Storry and Olsson, 2009). Finally, phenotypic classifications of weak subgroups do not always correlate with genetically characterized weak subtype alleles in publicly available databases of genetic variation, e.g., BGMUT (Yip et al., 2002; Yamamoto et al., 2014). This challenge in correlation of *ABO* genotype and ABO phenotype is a direct result of the vast amount of variation identified in *ABO* that converges on similar serological phenotypes.

1.5 ABO blood typing

ABO blood typing is the process of determining an individual's ABO phenotype. In theory it can be performed or imputed on multiple levels of biology by using multiple analytic modalities, among them, the genetic (DNA/mRNA), enzymatic (glycosyltransferase), glycoconjugate (carbohydrate antigen), and immunological levels (antibodies) (see Figure 1.6. for principles of the ABO system that are used to determine "type"). Historically there are two categories of ABO blood typing on red blood cells (1) forward typing (detection of carbohydrate antigens on the glycoconjugate level) and (2) reverse typing (detection of antibodies on the immunological level) (Storry and Olsson, 2009). Forward typing utilizes antibodies to detect A and B antigens on the surface of RBCs (Storry and Olsson, 2009). Currently these antibodies are routinely used to detect A and

B antigens on RBCs in clinical settings all over the world (Yamamoto et al., 2012). Recently potent anti-A antibodies have been developed with the goal of detecting weak A phenotypes such as the A subtypes A2, Ax, and many others (Yamamoto et al., 2012). Additionally, anti-B, and anti-H monoclonal antibodies have been produced and are commercially available in blood centers all over the world (Yamamoto et al., 2012). Conversely, for reverse typing, the detection of antibodies in the serum of an individual against the A or B antigen present on the surface of the RBC is performed using RBCs with a known blood type.

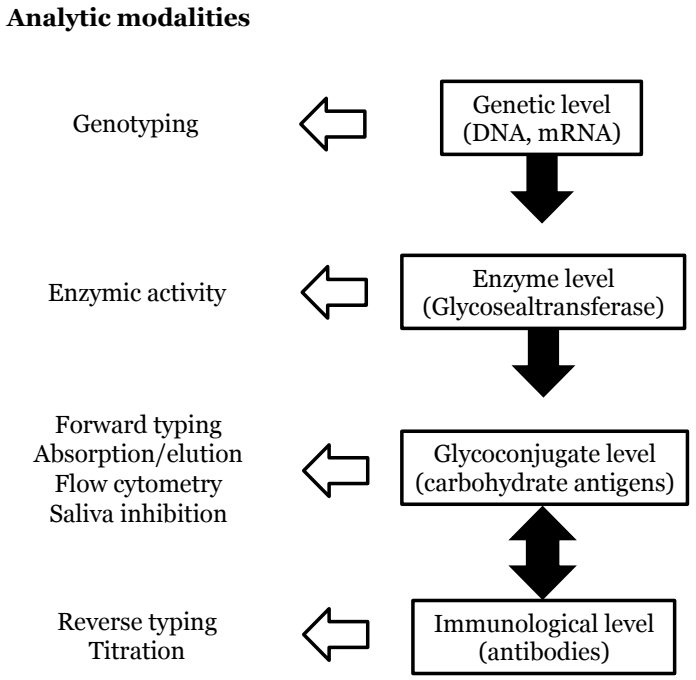

Figure 1.6: **The four principle levels used to determine *ABO* type.** Principles of the ABO typing at four molecular levels and interactions between these levels are schematically represented by arrows in black. On the left side, the analytical modalities currently used at each level are listed. (Adapted from Storry and Olsson, 2009.)

Lectins, carbohydrate-binding proteins derived from plants, are important tools in the identification of different blood groups (Khan et al., 2008). Lectins are any class of

molecules that bind in specific patterns to sugars and, as a result, have the potential as reagents to cause glycan-specific agglutination of particular cell types (Nilsson et al., 2007). They are routinely used to bind carbohydrate residues, including oligosaccharides such as the A antigen (Nilsson et al., 2007). DBA or *Dolcihos biflorus* agglutinin lectin is interacts with N-acetylgalactosamine and has been used to differentiate between A1 and A2 subgroup RBC's (Storry and Olsson, 2009). A lectin isolated from a West African shrub, *Griffonia simplicifolia,* can be used to bind to the B antigen, though it is less sensitive when binding B subtypes (Khan et al., 2008).

Advances in diagnostic technology have enabled the detection and quantification of antigens via alternative techniques that provide higher specificity and sensitivity than traditional forward-typing techniques. Flow cytometry is one of these tools. By suspending RBCs or platelets in a stream of fluid and passing them by an electronic detection apparatus, it can detect antigen presence and density with high sensitivity, including density at the single-cell level. Flow cytometry also allows for simultaneous multi-parametric analysis of physical and chemical characteristics, e.g., it can be used to simultaneously detect the density of both the A and H antigens on RBCs or platelets (Cooling et al., 2005). Flow cytometry is also able to differentiate between A1 and A2 antigen density on the surface of RBCs (see Figure 1.4) (Olsson et. al., 2010) and can be used to detect ABO blood-group antigens on organs and tissues (Yamamoto et al., 2014).

Within the field of immunology, *ABO* and the human leukocyte antigen (HLA) were the first loci where genotyping was applied to determine phenotype (Yamamoto et al., 2012). Yet *ABO* typing using genotype data is faced with some unique limitations. Although genetic assays are being used on a limited basis for other blood group genes, very few *ABO*-specific assays exist (Lane et al., 2016). There were several SNV-based assays for common variants in *ABO*, e.g., GTI Diagnostics previously distributed a 7-SNV *ABO* kit that reference laboratories occasionally employed, though it was not FDA approved (Jill Johnsen, pers. comm.). Because the majority of variation contributing to

critical amino acid substitutions is located in *ABO* exons 6 and 7, the majority of genotyping systems developed to determine ABO blood type only include a subset of variation, i.e., exons 6 and 7. They do not include non-coding/regulatory variation, which could produce aberrant messenger RNAs as a result of the formation of a new splice-site, or changes in expression levels as a result of variation located in either a promoter or enhancer (Yamamoto et al., 2014). These *a priori* SNV chip designs also exclude any other missense or LOF variation that potentially influence *ABO* glycosyltransferase activity.

Finally, because chip-based SNV approaches are limited to known, common, un-phased variation in the *ABO* gene, they are not capable of discovering rare variants that could have phenotypic consequences. The future of ABO clinical blood type testing likely includes a combination of both serological diagnostic technologies and genetic approaches that detect common and rare variation which might alter the structure and or function of the ABO glycosyltransferase (NGS or an equivalent technology) (see Chapter 5 for future directions and functional analysis).

1.6 ABO in clinical practice, transfusion therapy, and organ transplantation

More than 23 million units of blood are transfused each year in the USA (American Association of Blood Banks, 2008). Blood type compatibility matches have been successful in nearly eliminating major life-threatening transfusion reactions. However, adverse outcomes, including alloimmunization due to donor/recipient antigen mismatches, are still common in multiply transfused patients (Johnsen et al., 2016). Amongst the over thirty major blood group systems, the ABO blood-group system is one of the most clinically important blood group systems to consider in transfusion therapy and organ transplant compatibility (Yamamoto et al., 2014).

An FDA report in 2008 revealed that blood-group incompatibility accounted for 37% of all transfusion-related fatalities, with TRALI responsible for at least 35%

(www.fda.gov/cber/blood/fatalo8.htm). According to another FDA report on hemolytic fatalities, ABO was responsible for 59% of deaths (www.fda.gov/cber/blood/fatalo8.htm). An earlier study by Linden et al. (2000) found that 50% of patients who are inadvertently transfused with ABO-mismatched units of blood do not develop any signs of transfusion reactions. The mechanisms underlying the capacity for some patients to undergo ABO-mismatched transfusions and avoid transfusion reactions are not clearly understood. However, if this resistance mechanism can be understood, e.g., perhaps in the form of allelic variation, it could be exploited as a tool to increase organ transplantation success (Olsson and Storry, 2009).

Hemolytic disease of the fetus or newborn (HDFN) caused by ABO incompatibility can be relatively common, especially among O blood type mothers carrying an A or B blood type fetus (Olsson and Storry, 2009). However common, the effects of ABO incompatible HDFN are generally minor and seldom require treatment in a clinical setting (Olsson and Storry, 2009). In cases of anti-ABO mediated HDFN requiring treatment, the indication is often hyperbilirubinemia which can usually be managed with phototherapy.

The clinical significance of anti-A and anti-B antibodies has a much larger impact than transfusion therapy. ABO compatibility is also important for both solid organ transplants and hematopoietic transplantation (Olsson and Storry, 2009). Because of the limited supply of solid organs, there is intense interest in ABO-incompatible (ABOi) organ transplantation. Multiple contributions to the literature support the observations that children younger than three years old have been shown to tolerate ABOi organs better than adults. This is likely the result of relatively immature B-cell response among infant patients (West et al., 2001, West et al., 2016). ABOi transplantation is now established for kidney transplants with consideration for A subtypes and anti-ABO titers (Böhmig, et al. 2015), and the ABO allogeneic experience is expanding in hematopoietic

stem cell (Staley et al. 2016), liver (Kim et al. 2016), and heart (Urschel et al. 2016) transplants.

1.7 ABO and broader disease implications

ABO antigens are not limited to the cell surface of blood cells. They can also be found on a variety of human tissues, including the epithelium, sensory neurons, platelets, and the vascular epithelium (Yamamoto et al., 2012). Variation in ABO blood type is associated with disease susceptibility in cardiovascular disease, ischemic stroke, deep venous thrombosis, bleeding, various cancers, infectious diseases, keloid scarring, and ulcers (Yamamoto et al., 2012).

Although there is strong evidence ABO blood type plays a role in thrombotic vascular disease, less is known about the role of ABO haplotypes composed of both rare and common variants that manifest in specific subtypes. Moreover, almost nothing is known about the role of individually rare variants in the *ABO* gene and their role in disease susceptibility; this is a promising avenue for genomic-based *ABO* subtype determination. See Chapter 5 on future directions for more details on *ABO* haplotype structure and disease impact.

1.8 *ABO* genomics

Over the last decade, emerging genomic technologies have been developed to examine the genome in novel ways with unprecedented sensitivity, specificity, and throughput. The vast quantities of genetic data produced through public efforts in multiple human populations through the application of next-generation sequencing (NGS), offers potential for studying the number and types of rare variants that are present in sequenced human populations. Because of the importance of ABO in transfusion medicine and organ transplantation, and its possible role in common complex disease susceptibility, exploring these publicly available NGS datasets could provide new

insights and lead to the development of new tools to determine ABO blood types from sequence data.

Even though the *ABO* gene has already been mapped and common haplotypes are well characterized in multiple human populations, Bloodworks Northwest (formerly the Puget Sound Blood Center, a large regional blood bank in Seattle, WA) rarely uses genetic data to determine ABO blood type (Jill Johnsen, pers. comm.). One possible obstacle is that *ABO* polymorphic sites associated with antigen expression in the primary literature and reference databases, such as the International Society of Blood Transfusion (ISBT) and the BGMUT, are organized according to nucleotide positions in cDNA. This makes antigen prediction from next-generation sequencing data challenging, since it uses genomic coordinates (Lane et al., 2016). In fact, of the 56 contributing research groups responsible for cataloging 377 *ABO* subtype alleles (both intronic and exonic) in the NCBI's blood group gene mutation database (BGMUT), none has used a genomic-based approach to identify alleles. As a consequence, the alleles reported are in cDNA coordinates rather than genomic coordinates, for example, hg19 (http://www.ncbi.nlm.nih.gov/projects/gv/mhc/xslcgi.cgi).

Additionally, over the past five years multiple public efforts have been generated with NGS, including the 1000 Genome Project (1KG), and the NHLBI exome sequencing project, and many others (Auton et al., 2015). As a result, NGS-derived *ABO* genotype data are available for ~2,500 whole genomes, while allele frequencies for ~66,000 human exomes are all available in genomic coordinates. These NGS datasets provide a rich resource for cataloging potential common and rare variation in *ABO* in multiple human populations. Nevertheless, they are lacking corresponding ABO serological phenotype data.

In order to translate the information which can be gained from genomic approaches, two simple changes must occur. First, both serological phenotype data and NGS-sequence genotype data sharing need to be included in projects designed in the future

(Johnsen, el al., 2014). Second, there is enormous potential in mining and translating *ABO* alleles identified and then merging these with previous studies that include both serological data and accurate annotation of *ABO* alleles, even when different nomenclature systems are used (i.e., cDNA, BGMUT). These two approaches offer a new path forward for typing the complex human genetic variation in both the coding and noncoding portion of the *ABO* gene allowing for the development of sequence-based tools to create higher resolution cross-matches for patients who need transfused blood.

Since Yamamoto's contributions to ABO blood-group research through the description of the main ABO alleles (A, B, and O), many additional alleles have been characterized, annotated, and deposited in public databases such as the National Center for Biotechnology Information's (NCBI) Blood Group Gene Mutation database (BGMUT). In this dissertation, I mine existing human-genome sequence data in both the BGMUT and multiple-NGS datasets to develop a more comprehensive understanding of both coding and noncoding variation in the *ABO* gene in multiple human populations. My goal is to characterize both common and rare *ABO* haplotype structure (variants that are inherited together from one parent) in multiple human populations using whole genome, whole exome, and targeted-capture NGS datasets. I describe my efforts to construct a tool for imputing *ABO* subtype variation from *ABO* haplotypes assembled from NGS data, while simultaneously discovering novel *ABO* subtype variation from existing sequence-based algorithms. Finally, I explain how a more complete understanding of *ABO* haplotype structure in multiple populations could lead to (1) higher-resolution disease association studies, (2) predictive precision medicine using *ABO* as an example, and (3) the exploration of genetic variation in underrepresented minority populations, potentially establishing an opportunity to reduce the widening gap in health disparities.

The research summarized in Chapter 2 represents my investigation of *ABO* haplotype structure derived in multiple large-scale next-generation sequence (NGS) datasets.

Through these efforts two NGS datasets were investigated: (1) whole exome data through a combined dataset composed of the NHLBI-ESP and MH-GRID (n=6,432) and (2) whole genome data from the 1,000 Genomes Project, Phase 3 (n=2,504) which together total data from nearly 9,000 individuals from 26 different populations or (~18,000 *ABO* chromosomes), as well as two ancient hominid genomes. Through these efforts I applied a statistical phasing method (PHASE) to resolve *ABO* haplotype structure in the coding portion of the *ABO* gene in these datasets. Through our analysis we have identified the common variants known to influence ABO function: those known to the common A and B haplotype as well as a common (MAF >1) exon 6 indel that leads to the O genotype, and another common indel that results in the A2 subtype. We have also identified rare (MAF <1) coding variants within *ABO* (single nucleotide/missense variants, insertion/ deletions) that segregate on known haplotype backgrounds including rare loss of function (LOF) haplotypes such as the non-deletional O and Cis-AB haplotypes.

Chapter 3 presents my use of NGS data to increase the granularity at which the hematology community determines ABO subtype. Through these efforts, I have developed an algorithm (ABO-Seq) to determine the ABO blood type and subtype based on haplotypes characterized in the coding portion of the *ABO* gene within both the BGMUT and multiple NGS data-sets; these are derived from diverse genome sequencing platforms, including whole exome (the NHLBI-ESP and MH-GRID, n=6,432), and custom/targeted capture (the Bloodworks Northwest Blood-Seq Project, n=1,140). For 80 individuals within the NHLBI-ESP cohort (training set 1), and 469 individuals within the Bloodworks Northwest's Blood-Seq dataset (training set 2), clinical ABO serological analysis is available, which I used as a test set for matching with sequenced-based predictions to validate NGS based *ABO* subtyping.

In Chapter 4 I demonstrate the use of existing read-depth (RD) based methods to discover and validate, population-specific structural variation (SV) in the coding portion of *ABO* in a combined exome sequence data-set of 6,432 individuals, approximately half

of whom are of African ancestry and half are of European ancestry. I also determined the population frequency of this novel SV in an orthogonal dataset (the 1,000 Genomes Project, Phase 3) and provide a mechanism for the origin of this novel SV. Finally I prove that NGS read-depth-based methods have the sensitivity to discover novel *ABO* subtypes that would not be detectable using traditional serological methods.

In Chapter 5 I present future ideas for ABO blood-group genomics, including increased accuracy, automation, and the potential for widespread adoption of sequence-based screening methods in clinical settings. In particular, I describe how the combination of both serological and NGS-based systems (e.g., ABO-Seq) might allow clinicians to better manage transfusion compatibility and organ transplantation, subsequently refining our understanding of common complex disease association studies. I discuss the importance of functionalizing variation of unknown significance discovered in large-scale population-based screens. I also discuss the importance of characterizing the complete spectrum of human genetic variation and the importance of including underrepresented minority populations to achieve this goal. I conclude with some final thoughts on the future of genomic-based medicine.

(THIS PAGE LEFT INTENTIONALLY BLANK)

Chapter 2

INVESTIGATING NEXT-GENERATION SEQUENCE DATA ACROSS THE *ABO* LOCUS

Analysis of *ABO* haplotype structure derived from large-scale next-generation sequence (NGS) data holds promise for the diversity of this locus and can aid in the development of higher resolution ABO blood diagnostics (i.e., subtyping) in multiple populations. Two NGS datasets are investigated: (1) whole exome data through a combined dataset composed of the NHLBI-ESP and MH-GRID (n=6,432), and (2) whole genome data from the 1,000 Genomes Project, Phase 3 (n=2,504) which together total data from nearly 9,000 individuals, from 26 different populations or (~18,000 *ABO* chromosomes). Here we apply a statistical phasing method (PHASE) to resolve *ABO* haplotype structure in the coding portion of the *ABO* gene in these datasets. Through our analysis we have identified the common variants known to influence *ABO* function, including those known to the common A and B haplotype as well as a common (MAF >1) exon 6 indel that leads to the O genotype, and another common exon 7 indel that results in the A2 subtype. We have also identified rare (MAF <1) coding variants within *ABO* (single nucleotide/missense variants, insertion/deletions) that segregate on known haplotype backgrounds including rare loss of function (LOF) haplotypes such as the non-deletional O and Cis-AB haplotypes. These analyses are important for future studies of the locus to: (1) improve the specificity of ABO blood typing at both the clinical and research level by identifying rare functional alleles that might result in atypical serological patterns, (2) illuminate *ABO* gene architecture on a global scale, and (3) potentially identify novel associations between *ABO* and multiple human phenotypes.

2.1 Introduction

Human genomes are diploid and, for their complete description and interpretation, it is necessary not only to discover the variation they contain but also to arrange it onto chromosomal haplotypes. Although sequence data is becoming increasingly routine, nearly all such individual genomes are mostly unresolved with respect to haplotype, particularly for rare alleles, which remain poorly resolved by inferential methods (Snyder, Adey, Kitzman, and Shendure, 2015). Haplotypes are in fact aggregated variation inherited from a single parent. This aggregated variation (for our purposes multiple alleles) is

likely to be conserved as a sequence that survives the descent of many generations of reproduction (Snyder et al., 2015). This aggregate of variation is inherited together because of genetic linkage (Morgan, 1904), or the phenomenon by which genes that are close to each other on the same chromosome are often inherited together on haplotype "blocks." Genetically older populations (e.g., Yoruaba of Nigeria) tend to have shorted haplotype blocks with respect to younger populations (e.g., Europeans of Utah) because older populations have accumulated a higher number of recombination events over generational time. This is one reason why older populations (e.g., sub-Saharan Africa) have accumulated more genetic diversity over time.

Regarding *ABO* haplotype structure there are a few rules that govern our observation of human genetic variation. ABO is a human locus that has undergone selective pressure maintaining the observed variation in *ABO* haplotype structure among human populations (i.e., A, B, and O) is an extraordinary yet classic result of balancing selection on the human genome. The second is that while *ABO* haplotypes have a common progenitor haplotype (i.e., common aggregated variation), rare variation (MAF <1%) can accumulate on common haplotype backgrounds and eventually alter the structure/ function of the glycoseyltransferase giving rise to either (1) loss of function variation (e.g., A -> O haplotype) or (2) change of function (e.g, A -> B haplotype). Finally, this rare variation is usually population-specific and recently explored on a global scale *via* large-scale population-based studies (e.g., 1,000 Genome Project).

2.1.1 Distribution of ABO blood types among human populations

Hirschfeld, who described ABO as an inherited phenotype, was the first to conduct a study of population based on a genetic survey of blood types which was carried out during the First World War (Hirschfeld and Hirschfeld, 1919). They were not only the first to connect the observation of variation in ABO blood type serology as an inherited genetic phenotype but also the first to suggest that the genetic variation observed in ABO

might provide insight into understanding phenotypic differences in human populations. To quote Hirschfeld and Hirschfeld, "it was clear to us from the beginning that we could only attack the human race problem on serological lines" by using the "iso-agglutinins first analyzed by Landsteiner." By "the human race problem" he meant an understanding the population frequencies in A, B, and O blood types among diverse populations (Hirschfeld and Hirschfeld, 1919). In this study, they attempted to establish ABO blood type frequencies among soldiers of various ethnicities in the Royal British Military by examining soldiers' blood. These early experiments would yield the first population-specific estimates of A, B, and O blood types, but more importantly the birth of a sub-discipline within human genetics: population genetics. However, it wasn't until 1924 when Felix Bernstein applied a simple proof, the foundation for modern diploid population genetics, then called the "Hardy-Weinberg Law" to generate accurate allele estimates demonstrating that the variation observed in ABO blood type among human populations was not found in multiple genes, but multiple alleles in one gene, that were necessary to generate the correct frequencies of blood types observed in multiple human populations (Crow, 1993).

By the 1960's *ABO* "gene frequencies" had been categorized for many human populations. Dr. Luca Cavalli-Sforza recognized that in order to study living populations it would be useful to create "geographic representations" of blood type gene data for aboriginal populations. Cavali-Sforza noted in his seminal text, *History and Geography of Human Genes*, that allele frequencies alone are "inadequate" when attempting to determine human evolutionary history, and that: (1) gene frequency varies over time in ways that can be considered, at least superficially, nearly random and (2) therefore, it is not surprising that populations having clearly different evolutionary histories may show similar gene frequencies. However this drawback can be avoided if one "cumulates" the information from more than one gene. In 1963 using this multivariate approach Cavalli-Sforza and Edwards demonstrated that even with as few as 20 alleles from five genes one could successfully create a reconstruction of human evolutionary history (Cavalli-Sforza,

Barrai, and Edwards, 1964). The *ABO* locus was one of those five genes used in these analyses.

Cavalli-Sforza would use this "geographic" perspective to create some of the first allele frequency heat maps of *ABO* among aboriginal populations (Mountain and Cavalli-Sforza, 1994). Interestingly Cavalli-Sforza's early *ABO* heatmaps indicated that the B blood type, which is the rarest blood type of the three commonly observed blood types (A, B, and O, with the exclusion of the AB blood type), is highest amongst both South and East Asian populations. The frequency of B blood type is especially high in India, supporting Hirschfeld's original estimates calculated among Indian soldiers enrolled in the British military during World War I (Hirschfeld and Hirschfeld, 1919).

In the 1976 Arthur Mourant published *The Distribution of the Human Blood Groups and Other Polymorphisms* that revealed interesting global "clines" or "gradations" in ABO blood type distribution (Godber et al., 1976). According to Mourant (1976), while type O was observed at the highest frequency globally at over 60%, as populations moved east and north through present day Russia and into both North and South America, the percentage of the O phenotype increased from around 60% in what is now Russia to greater than 80% in North America and finally reaching greater that 90% in South America (Figure 2.1) (Villanea et al., 2013). Deborah Bolnick et al., at the University of Texas, Austin believes that the reason the O blood group nearly reaches fixation amongst many contemporary Native American populations is in part due to the selective pressure of being exposed to smallpox via initial European contact (Halverson and Bolnick, 2008).

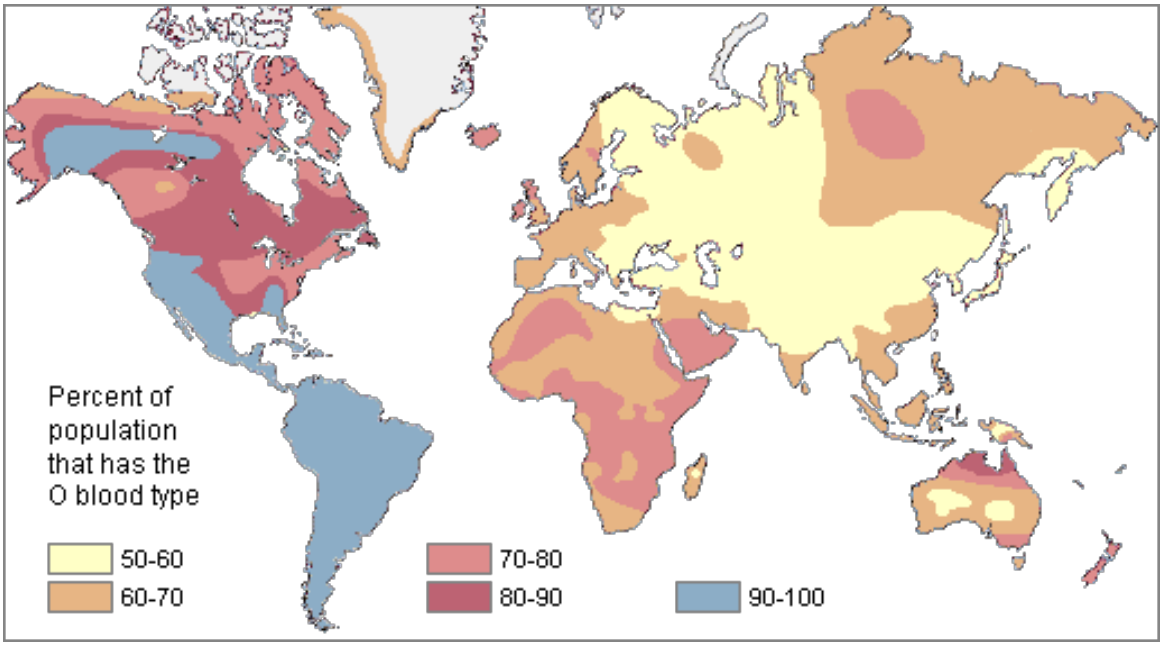

Figure 2.1: **Blood types vary depending on the geographical region (the O blood group).** The O blood type is the most common blood type around the world, and is carried by nearly 100% of those living in South America. It is the most common blood type among Australian Aborigines, Celts, those living in Western Europe, and in the United States. Scandinavians have a high probability of carrying the A blood type, while those indigenous to central Asia are more likely to carry the B blood type. (Modified from Garratty et al., 2004.)

Finally, historical approaches from Cavalli-Sforza and Mordant highlight the utilization of serological-based determination of ABO blood type combined with early statistical approaches (i.e., Hardy-Weinberg disequilibrium) to determine *ABO* allele frequencies accurately recapitulating human migration patterns (Cavalli-Sforza et al., 1964). However, calculating allele frequencies based on serological phenotype alone as an estimator of genotype frequencies is not sensitive enough to detect the effects of rare variation that occasionally manifest in loss of function (LOF) haplotypes, unique Cis-AB haplotypes and detection of population specific rare variation that has the potential to alter common A, B, or O haplotype backgrounds. While common haplotypes are responsible for the lion's share of diversity in ABO, rare haplotypes exist and can be

utilized as a tool to understand population migration history as well as natural selection in this locus due to historical interactions with pathogens (e.g., malaria, norovirus, smallpox, and other unidentified pathogens).

2.1.2 Common ABO haplotype structure

In addition to the classic LOF frameshift discovered by Yamamoto et al. in *ABO* exon 6 responsible for the O haplotype, and a read through frame shift variant at the end *of ABO* exon 7 that leads to A2, four common missense variants are crucial for determining ABO blood type and thus serve as the backbone for defining common A and B haplotypes. As depicted in Figure 2.2 the majority of known common haplotype variation occurs in exons 6 and 7 of the *ABO* gene. The four common nucleotide substitutions (missense variants) in the coding sequence of *ABO* exon 7 reported to differentiate between the A and B haplotype result in amino acid substitutions in the active/binding site of the *ABO* glycosyltransferase: rs7853989 (p. R176G), rs8176743 (p. G235S), rs8176746 (p. L266M), and rs8176747 (p. G286A) (Yamamoto, 1992).

Exon Number		6	7															
Nucleotide position		261	297	467	526	646	657	681	703	771	796	802	803	829	871	930	1054	1060
A alleles	A101	G	A	C	C	T	C	G	G	C	C	G	G	G	G	G	C	C
	A201	*	*	T	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ
B alleles	B101	*	G	*	G	*	T	*	A	*	A	*	C	*	*	A	*	*
O alleles	O01	Δ	*	T	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Possible amino acid change		Frameshift	No change	P156L	R176G	F216I	No change	No change	G235S	No change	L266M	G268R	G268A	V277M	D291N	No change	R352W	Frameshift

Figure 2.2: **Common *ABO* variants and their corresponding amino acid substitutions.** Some of the common haplotypes in the *ABO* locus. Adapted from (Yamamoto et al., 2014).

Although these four variants define the majority of common A and B haplotype background variation that exists in the *ABO* gene, the effects of rare variation in *ABO* has yet to be fully explored. Rather, variation in *ABO* should be thought of as a combination of both common (MAF <1%) and rare variation (MAF >1%) resulting in altered *ABO* glycosyltransferase efficiency and specificity (Patenaude et al., 2002). Moreover, more recent rare population-specific SNVs have accumulated on older common haplotype backgrounds leading to more complex ABO haplotype structure. However, the majority of population based screens of genetic variation in the *ABO* gene have primarily included individuals of European Ancestry (Bustamante, Burchard, and la Vega, 2011), and thus little is known about the effect of rare population specific variation in *ABO*.

The application of massively parallel sequencing is now providing diverse large-scale datasets to characterize the diversity of *ABO* on a global scale. Here, I combine three datasets from the the NHLBI Exome Sequencing Project (NHLBI-ESP) (Auer et al., 2012) — Minority Health GRID (MH-GRID) (Seffens, Evans, Minority Health-GRID Network, and Taylor, 2015), and the 1,000 Genomes Project — to characterize human genetic diversity in the form of common (MAF <1%) and rare (MAF >1%) variation (e.g., single nucleotide variants [SNVs] insertion/deletion (indel) variants, and in another chapter structural variants [SVs]) and statistical phase these variants *(*SNVs and indels) to resolve *ABO* haplotypes composed of both common and rare variation. Finally, I compare those imputed *ABO* subtypes (paired haplotypes) in 2,504 individuals from five distinct continental populations (i.e., the 1,000 Genomes Project) to serologically derived ABO blood types observed across the globe.

2.2 Methods and subjects

2.2.1 The next-generation datasets used for ABO analysis

I explore variation in NGS datasets generated obtained using two distinct sequencing approaches: (1) whole exome and (2) whole genome.

2.2.2 Aggregating multiple human exome sequencing datasets

We used *ABO* coding sequence data derived from the NHLBI Exome Sequencing Project (ESP) and the Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID). Through the ESP, 15,336 genes were sequenced at high coverage (median depth >100x) in a total of 6,515 unrelated European American (EA, n=4,298) and African American (AA, n=2,217) individuals from 19 different cohorts. Through MH-GRID, exome sequences at similar high coverage (median depth >100x) were obtained from a total of 1,313 unrelated AA individuals. The library construction, exome capture, sequencing, mapping, calling, and filtering were carried out as described previously (Auer et al., 2012). Exome sequence data were aligned to NCBI human reference GRCh37.

Sequence data were aligned to reference sequence obtained from GenBank (NG_006669.1). Average sample read depth for the *ABO* gene was 77x (ranging from 10 to 374x) and includes the entire coding sequence and exon-intron boundaries (see Table 2.1). Of the 4,298 EA ESP participants, 3,405 had minimum coverage of 50x across the entire targeted ABO region. Additionally, 3,027 self-identified African Americans (AA) from both the ESP and MH-GRID were included in our ABO analysis bringing our total to 6,432 participants, roughly equally divided between EA (53%) and AA (47%). Finally, linking NGS derived genotype to ABO serology for 80 individuals included in the ESP dataset (see Table 2.1).

2.2.3 An NGS Global reference panel for human genetic variation

Sequence data from ABO (5,008 *ABO* chromosomes) were also obtained from a publicly available global reference panel, the 1,000 Genomes project, which contains a diverse set of individuals from multiple populations. Through the 1,000 Genomes project, whole genomes of 2,504 individuals from 26 different populations were sequenced at low coverage (~6x). Those 26 populations (see Table 2.1) have been lumped into five

super-populations based on continent of origin: African (AFR), American (AMR), East Asian (EAS), South Asian (SAS), and European (EUR). The library construction, sequencing, mapping, variant calling, and filtering were carried out as described previously (Auton et al., 2015). It is important to note that each population included in the 1,000 Genomes project is from a geographically distinct population with four grandparents from that region. Finally, none of the individuals included in the 1000 Genomes Project had serological phenotype data to accompany their low-coverage derived genotype data.

Table 2.1: **Summary dataset for *ABO*.** Variants identified in both datasets were limited to coding variation derived from VCFs generated using called using GATK (McKenna et al., 2011).

Dataset	NGS platform	Depth	Samples	Ethnicity	Missense	Indels	Splice
NHLBI-ESP + MH-GRID	exome	~100x	6,432	2	46	7	2
1,000 Genomes Project	whole genome	~6x	2,504	26	44	4	2

2.2.4 Haplotype construction and determination of ABO subtype from the sequence datasets

We obtained genotype calls for the *ABO* locus from each NGS dataset included in our analysis in variant call format (VCF) file using the Genome Analysis Tool Kit (GATK) to call single nucleotide variants (SNVs) as well as insertions and deletions (indels) (McKenna et al., 2010). Both indels and SNVs are important for blood group calling because the primary differences between the A and B haplotypes are SNVs while the common cause of the O blood type is a single base deletion that causes nonsense mediated decay of the RNA transcript resulting in absence of protein (Yamamoto et al., 2012). In order to resolve *ABO* haplotypes from these NGS datasets, I employed PHASE 2.1.2 for haplotype construction of the different chromosomal alleles for each individual (Scheet and Stephens, 2006).

2.3 Results

2.3.1 Overview of variation in 17,872 *ABO* chromosomes

In a combined exome dataset consisting of samples from both the NHLBI-ESP and the MH-GRID, we investigated *ABO* variation in 6,432 diploid participants (i.e., the coding sequence of 12,864 *ABO* genes). The data were stratified by genetically determined ancestry using a principal components analysis. Among the exome samples, 3,405 samples are of European ancestry (EA) and 3,027 are of African Ancestry (AA). Of the 75 non-synonymous variants identified; 16 were common (MAF <1%), 6 had an MAF between 1-10%, and 53 were infrequent, or rare, with an MAF >1%. 12 of these rare variants were singletons, 4 were doubletons, and the remaining 37 had an MAF of >1%. A subset of these *ABO* variants potentially affect the structure and/or function of *ABO* glycosyltransferase. These non-synonymous variants categorically include missense variants resulting in potential amino acid substitutions (n=45), nonsense/frame shifts (n=7), and splice site variation (n=2).

Through the 1,000 Genomes Project, we investigated *ABO* variation in 2,504 diploid participants (i.e., the coding sequence of 5,008 *ABO* genes). The participants included in this dataset were from 26 distinct ethnic groups from 4 continents. Of the 68 SNVs identified; 16 were common (MAF <1%), 7 had an MAF between 1-10%, and 45 were infrequent, or rare, with an MAF >1%. 20 of these rare variants were singletons, 4 were doubletons, and the remaining 21 had an MAF of >1%. A subset of these *ABO* variants potentially affect the structure and/or function of *ABO* glycosyltransferase. These non-synonymous variants categorically include missense variants resulting in amino acid substitutions (n=44), nonsense/frame shifts (n=4), and splice site variation (n=2). The full spectrum of human genetic variation identified in the 1,000 Genomes dataset has been summarized previously in multiple publications (Sudmant et al., 2015).

2.3.2 Common missense variation in 17,872 chromosomes

Of the common (MAF >1%) missense and nonsense variation identified through our analysis of the multiple NGS datasets included in this study (Supplemental Table 2.1 and 2.3), 6 common missense variants are crucial for determining ABO blood type. As depicted in Figure 2.2, these common haplotype variants occur in exons 6 and 7 of the *ABO* gene (Yamamoto, McNeill, and Hakomori, 1995). These four common nucleotide substitutions in the coding sequence can be used to differentiate the majority of A and B haplotypes resulting in four amino acid substitutions located in the active/binding site of the *ABO* glycosyltransferase: rs7853989 (p. R176G), rs8176743 (p. G235S), rs8176746 (p. L266M), and rs8176747 (p. G286A) (Yip, 2002). Rare or infrequent SNVs (MAF <1%) accumulate on these common haplotype backgrounds (see Figure 2.4) resulting in more complex *ABO* haplotype variation (International HapMap Consortium et al., 2007).

2.3.3 Rare or infrequent missense variation in 17,872 *ABO* chromosomes

In addition to common nucleotide substitutions found in the active/binding site of the *ABO* glycosyltransferases, we identified novel, rare, or infrequent variation within several amino acids of the active/binding site of the *ABO* glycosyltransferase that might result in novel forms of A or B haplotypes. Identifying putatively functional variation in the coding region of *ABO* through prediction algorithms is challenging and these approaches have limitations, as a result many of the rare variants identified in large scale population based screens are considered variants with unknown significance (VUS) (Starita et al., 2015). However, many variants that effect the structure/function of the ABO glycosyltransferase are located at or near the active site of the glycosyltransferase (i.e., rs7853989 (p. R176G), rs8176743 (p. G235S), rs8176746 (p. L266M), and rs8176747 (p. G286A). For example, rs370138477 (p. E297K) — a singleton identified in our combined exome sequencing dataset located in a highly conserved position (based on GERP and polyphen2 scores) in exon 7 — represents a promising candidate for

follow-up functional testing and the potential identification of a novel *ABO* subtype in the future. For a complete list of rare or infrequent missense variants found in both datasets see Supplemental Tables 2.1, 2.2, and 2.3.

We were also able to identify non-deletional O (non-del O) and Cis-AB haplotypes defined by missense variants that result in amino acid changes in the 4 critical active site loci themselves (Patnaik, Helmberg, and Blumenfeld, 2014). The non-del O haplotypes identified — chr9:136,131,592 (rs7853989) — resemble the B haplotype (i.e., a C allele at one position), while the remaining active-site motif (rs8176743, rs8176746, and rs8176747) is that of an A haplotype (C, G, C, respectively). An isolated case of the Cis-AB haplotype was also discovered in an African Caribbean individual for Barbados identified in the 1,000 Genomes project dataset (n=1); its active site motif also resembled a combination of both the A and B haplotypes resulting in its predicted phenotypic affinity being able to attach both A and B sugars to the H antigen (Yip, 2002).

2.3.4 Indels found in 17,872 *ABO* chromosomes

Determination of ABO haplotype structure requires indel calling because the most common functional variant distinguishing O from A or B is an indel. There are two common single base-pair deletions that are necessary for the accurate determination of ABO blood type at the subtype or weak hemagglutination level. Using the genome analysis tool-kit (GATK) we were able to call variation and annotate both NGS datasets separately and successfully identify both common indels known to influence structure and function of the *ABO* glycosyltransferase. The well-known exon-6 deletion leads to the classic O genotype and phenotype when present in individuals who are homozygous for this variant (rs8176719, deletion located in exon 6 9:136132908) which usually appears on the A ancestral haplotype. Secondly, a common deletion located at the end terminus of exon 7 results in a frame shift read-through and the addition of +21 amino acids resulting in the A2 subtype (rs56392308, deletion, end terminus of exon 7) (Yamamoto, McNeill, and Hakomori, 1992). We were also able to identify less frequent

or rare indels that might affect structure and function of the *ABO* glycosyltransferase. In our combined exome dataset we identified a premature stop codon located in the beginning of exon 7, downstream from the classic O indel (rs56284703, deletion, MAF = .22%, exclusively found in individuals of European ancestry). In the absence of the classic O deletion this potential loss of function deletion could also result in truncation of the *ABO* glycosyltransferase similar to the classic O indel (rs8176719). For a complete list of indels in both NGS datasets see Supplemental Tables 2.1, 2.2, and 2.3.

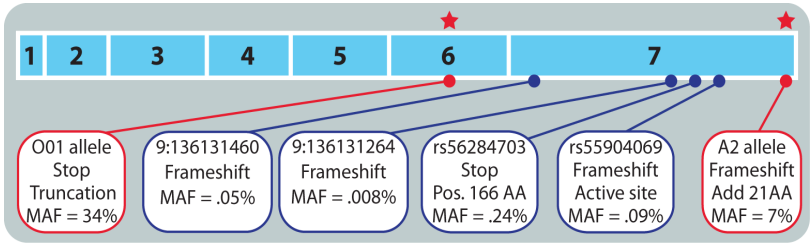

Figure 2.3: **Examples of known and novel predicted LOF alleles in the *ABO* gene.** AA = amino acid, MAF = minor allele frequency.

2.3.5 Splice site variants found in 17,872 *ABO* chromosomes

Although not traditionally associated with *ABO* subtype variation, alternative splicing could be responsible for variation we observe in the ABO blood type. We observed two five-prime splice site variants in both our combined exome dataset and the 1,000 Genomes Project (rs368673814 and rs28385716). According to the bioinformatics tool Human Splice Finder (http://www.umd.be/HSF/), rs368673814 is likely a canonical donor site mutation at the end terminus of exon 6 resulting in a premature stop codon. This potential LOF frame shift caused by a splice site variant might appear as O when ABO typed *via* hemagglutination, but the cause of ABO loss of function in this case would be the result of alternative splicing. This variant is an interesting candidate for molecular validation in the future.

2.3.6 Comparing ABO rare variation in both datasets

When comparing the rare variation found in our combined exome sequencing dataset to the variation found in the 1,000 Genome Project, we identified 44 coding variants that are found in both datasets, 27 rare variants that are exclusively found in our combined exome sequencing dataset, and 24 rare variants that were exclusively found in the 1,000 Genome Project dataset. When coming each dataset to the BGMUT ABO allele database (n=90 alleles total), we found that 29 variants discovered in our combined exome sequence data set were also allele entries on BGMUT, 41 were rare variants exclusively found in in our combined exome sequence dataset, and 59 allele entries existed in the BGMUT database that were not present in our combined exome sequencing dataset. When combing each of the 1,000 Genomes Project datasets to the BGMUT ABO allele database (n=90 alleles total), we found that 30 variants discovered in the 1000 Genomes dataset were also allele entries on BGMUT, 35 were rare variants exclusively found in the 1000 Genome Project dataset, and 56 allele entries existed in the BGMUT database that were not present in 1,000 Genome Project dataset. (For a summary of this comparison see Supplemental Table 2.3.)

2.3.7 High-throughput exploration of *ABO* haplotype diversity

ABO is co-dominant between the A and B phenotypes (e.g., the AB phenotype) and A and B can mask the observation of O because O is recessive to both A and B such that an individual that is A could have either haplotype combination, AA or AO. Similarly, an individual with the B phenotype could be BB or BO. This gene co-dominance makes it critical to differentiate between variation on one copy of *ABO* versus the other (i.e., multiple variants on one copy of the *ABO* gene, or haplotypes) via phasing to determine the genetic structure of the locus and its corresponding phenotype.

ABO haplotypes are composed of both common and rare variation in both datasets. Each haplotype in our combined exome sequence dataset is composed of 85 variable loci.

In our combined exome dataset of 6,432 individuals we characterized 211 unique *ABO* haplotypes from 12,864 *ABO* genes (Figure 2.4, Table 2.2). 114 of the 211 (54%) haplotypes discovered were singletons. 23 of the 211 were doubletons (11%), and 16 of the 211 were tripletons (8%). Of the 211 unique *ABO* haplotypes 135 (64%), were included the classic O indel (rs8176719, deletion). When characterizing common haplotype backgrounds, 54 (26%) included the common A haplotype background, and 22 (10%) included the common B haplotype background. Of the 12,864 *ABO* genes exome sequenced, 8,610 were classified as O haplotypes, 2,934 were classified as A haplotypes, and 1,320 were classified as B haplotypes. Within the common A haplotype group, 19 of the 54 (34%) included the A2 subtype indel (rs56392308, deletion).

In the 1,000 Genomes Project we characterized 108 unique *ABO* haplotypes from 5,008 *ABO* genes (Table 2.2, Figure 2.5a). Each haplotype is composed of 69 variable locations. Fifty-one of the 108 (47%) haplotypes discovered were singletons. 8 of the 108 were doubletons (7.4%), and 7 of the 108 were tripletons (6.4%). Of the 108 unique *ABO* haplotypes 72, or 66.6% included the classic O indel (rs8176719, deletion). When characterizing common haplotype backgrounds, 20 or 18.5% included the common A haplotype background, and 9 or 8.3% included the common B haplotype background. Of the 12,864 *ABO* genes exome sequenced 3,289 were classified as O haplotypes, 943 were classified as A haplotypes, and 719 were classified as B haplotypes. Within the A haplotype group, 8 of the 20 or 40% included the A2 subtype indel (rs56392308, deletion).

Table 2.2: **Summary of *ABO* haplotypes in both our combined exome dataset and the 1,000 Genome Project**. The summary of *ABO* haplotypes was constructed using PHASE 2.1.2. We have broken down the 1,000 Genomes Project into super-populations (admixed Americans, Africans, East Asian, European, and South Asian). Each cohort includes unique haplotypes that are then further broken down into A, A weak, B, O, non-deletional O, and Cis-AB haplotypes. The number of chromosomes identified in each haplotype category are shown in parentheses.

Cohort	Unique *ABO* haplotype	A haplotype	A weak	B haplotype	O haplotype	Non-del O	Cis-AB
ESP + MH-GRID	211	29 (1,958)	19 (821)	22 (1,320)	135 (8,610)	6 (155)	0
1,000 Genomes	108	12 (701)	8 (242)	9 (719)	72 (3,289)	6 (56)	0
Admixed American	31	3 (86)	1 (30)	1 (33)	28 (554)	2 (12)	0
African	58	4 (115)	5 (87)	2 (178)	42 (940)	3 (5)	1 (1)
East Asian	25	4 (151)	0	6 (198)	15 (659)	1 (1)	0
European	34	4 (185)	4 (96)	1 (85)	21 (640)	3 (30)	0
South Asian	28	3 (163)	3 (29)	3 (228)	16 (589)	3 (9)	0

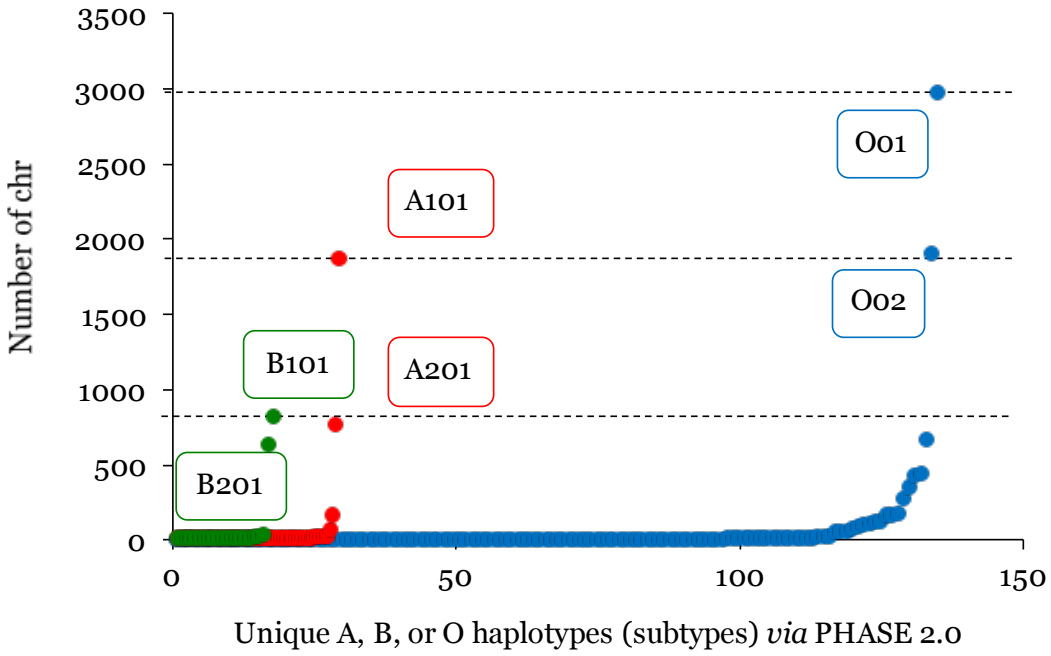

Figure 2.4: **Distribution of *ABO* haplotypes in a combined exome dataset of 6,432 individuals (12,864 chromosomes).** On the X-axis we have plotted the number of unique A (red), B (green), and O (blue), haplotypes (n=211 total) in our combined exome dataset of 6,432 individuals. On the Y-axis we have plotted the number of *ABO* chromosomes assigned to each haplotype A (n=), B (n=), or O (n=) (n=12,864 total). The majority of phased ABO genes are common haplotype forms of A, B, and/or O. However, the majority of the unique ABO haplotypes are original A, B, and/or O subtypes with one-off singleton variation on common A, B, or O backgrounds. Common haplotype subtypes (A101, A201, B101, B201, O01, and O02) were assigned via BGMUT variation look-up tables. It is important to note that visually these singletons are overlapping starting for 0-211.

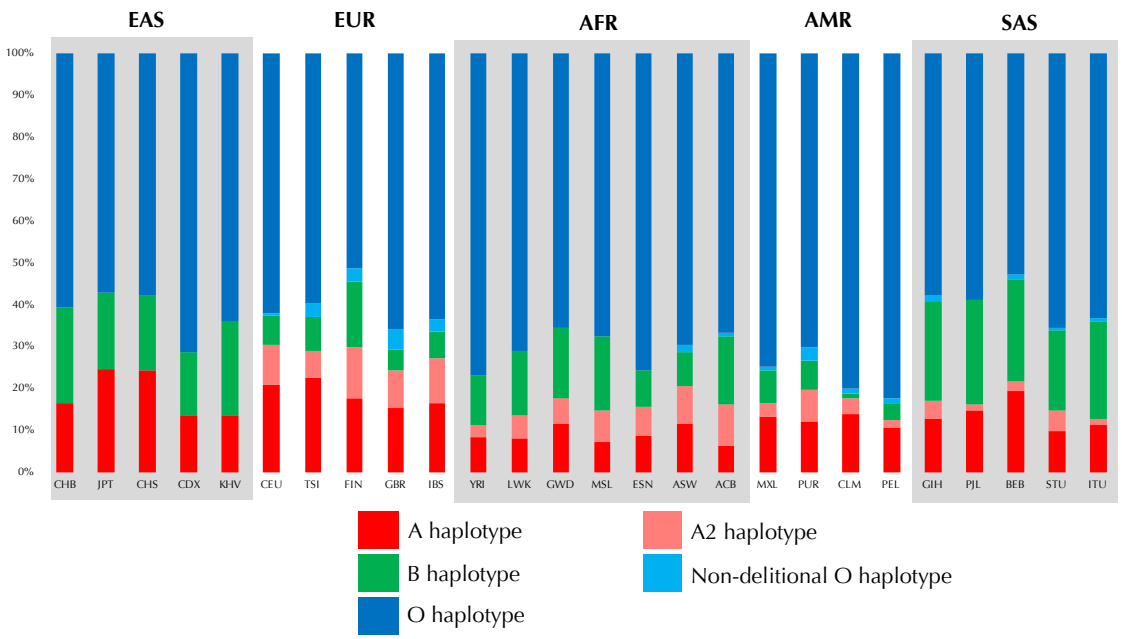

Figure 2.5a: ***ABO* haplotype diversity in the 1000 Genomes Projects dataset.** Stacked bar plots show the percentage of each ABO haplotype out of 100% for each of the 26 populations included in the 1,000 Genome Project. All super populations are represented above in these categories: admixed American (AMR), African (AFR), East Asian (EAS), European (EUR), and South Asian (SAS). Each population group is composed of almost 100 individuals; each super-population group is composed of approximately 500 individuals. ABO haplotypes were constructed using PHASE 2.1.2. and defined based on both the O indel, A2 indel, and 4 missense active-site variations. The A haplotype is shown in red, the B haplotype is shown in green, the O blood type is shown in blue, the A2 haplotype is shown in pink, and the non-deletional O haplotype is shown in light blue. For a complete table of blood types and haplotype counts divided into the 26 specific ethnic groups included in the 1,000 Genomes Project please see Supplemental Table 2.5.

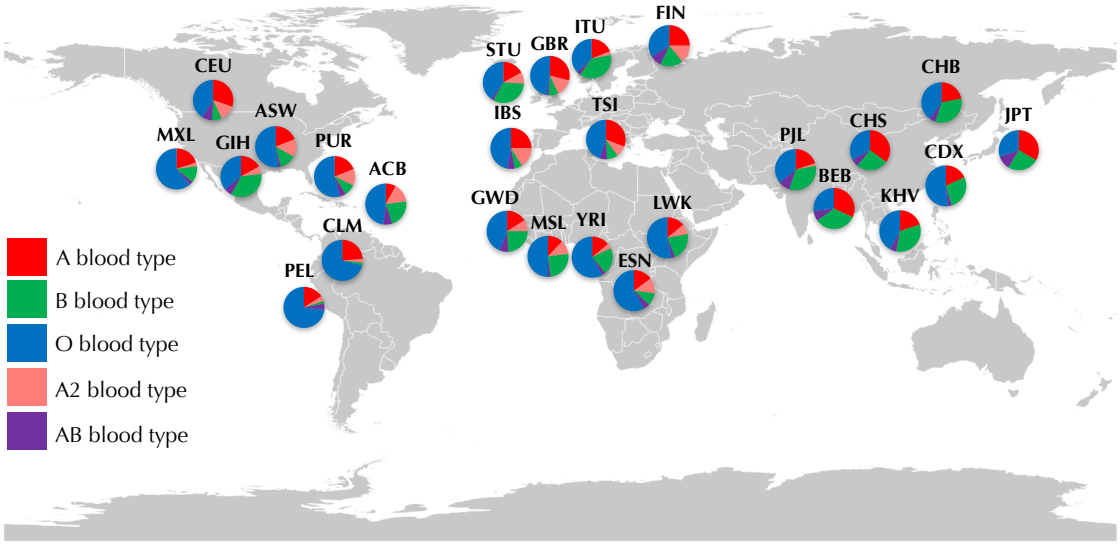

Figure 2.5b: **NGS Imputed ABO blood types for the 1,000 Genomes Project**. A map of the world highlighting the geographic location of all 26 populations in the 1000 Genomes Project (longitude/latitude coordinates were collected for each population for accurate geographic placement on the globe). ABO blood types were imputed by pairing and combining haplotypes constructed using PHASE 2.1.2. Each population group is composed of approximately 100 individuals. The A blood type is shown in red, the B blood type is shown in green, the O blood type is shown in blue, the A2 blood type is shown in pink, and the AB blood type is shown in purple. For a complete table of blood types and haplotype counts divided into the 26 specific ethnic groups included in the 1,000 Genomes Project see Supplemental Table 2.5.

Table 2.3: ***ABO* singleton SNVs identified in our combined exome dataset**. Eight singletons were identified in total. Seven of those 8 singletons (87.5%) were identified on O haplotype backgrounds. The remaining singleton haplotype was identified on a classic B haplotype background. This individual is of European ancestry.

Chromosome	HG19 position	Allele	Individual	Ethnicity	Haplotype Background
9	136131136	A	pcath981638-1	African Ancestry	O
9	136131190	C	3294	African Ancestry	O
9	136131264	TG	A06878	European Ancestry	O
9	136131268	T	77939	African Ancestry	O
9	136131563	A	74474	African Ancestry	O
9	136132867	T	258571-49	European Ancestry	B
9	136135260	C	332278	African Ancestry	O
9	136137521	T	3371	African Ancestry	O

Table 2.4: ***ABO* singleton SNVs identified in the 1,000 Genome Project.** Twenty singletons were identified in total. Eighteen of those 20 singletons (90%) were identified on O haplotype backgrounds. The remaining 2 haplotypes (10%) were identified on classic A haplotype backgrounds. Both of those individuals with the A singleton haplotypes were of Han Chinese ancestry.

Chromo -some	HG19 Position	Allele	Individual	Ethnicity	Haplotype Background
9	136131059	C	HG02792	Punjabi from Lahore, Pakistan	O
9	136131065	A	NA19788	Mexican Ancestry from Los Angeles USA	O
9	136131086	T	NA20772	Toscani in Italia	O
9	136131118	T	NA20351	Americans of African Ancestry in SW USA	O
9	136131154	T	NA18639	Han Chinese in Bejing, China	A
9	136131240	C	HG02623	Gambian in Western Division, Gambia	O
9	136131319	T	HG03762	Punjabi from Lahore, Pakistan	O
9	136131556	T	NA18605	Han Chinese in Bejing, China	A
9	136131616	C	NA20505	Toscani in Italia	O
9	136131630	A	HG00142	British in England and Scotland	O
9	136131664	G	NA18941	Japanese in Tokyo, Japan	O
9	136131704	T	HG02697	Punjabi from Lahore, Pakistan	O
9	136131718	A	HG01773	Iberian Population in Spain	O
9	136132845	G	HG01676	Iberian Population in Spain	O
9	136132853	C	NA19771	Mexican Ancestry from Los Angeles USA	O
9	136132864	A	NA18961	Japanese in Tokyo, Japan	O
9	136135226	C	HG03713	Indian Telugu from the UK	O
9	136135232	A	HG02881	Gambian in Western Division, Gambia	O
9	136136728	G	HG01951	Peruvians from Lima, Peru	O
9	136137551	A	HG02181	Chinese Dai in Xishuangbanna, China	O

2.3.8 Ancient hominid *ABO* haplotype structure

In addition to analyzing *ABO* haplotype structure, in both our combined exome sequencing dataset and the 1,000 Genomes Project, we extracted coding variation from two ancient hominid VCFs pertaining to a Neanderthal individual from Siberia (~42X coverage for the coding region of *ABO*) and a Denisovan individual from the Altai Mountains (~21X coverage for the coding region of *ABO*) (Prüfer et al., 2014; Meyer et al., 2012). Both of these individuals are estimated to be at least 50,000 years old. To include this dataset, we extracted all 69 loci identified in our initial phasing of the 1,000 Genomes Project and re-phased the 1,000 Genomes Project dataset including both ancient hominids (n=2,506 individuals). This resulted in a total of 111 unique haplotypes. Both the Neanderthal and Denisovan had distinct *ABO* haplotypes not found in any contemporary human population (see Supplemental Table 2.4).

After pairing haplotypes both ancient hominids typed as O. The Denisovan individual had two unique forms of deletional O not found in contemporary populations (Figure 2.6). Both haplotypes for the Denisovan individual were singletons and share a rare missense mutation in amino acid 163 (9:136131630). Upon querying this locus on ExAC it is rare and only found in European (allele frequency = 0.0008893), South Asian (allele frequency = 0.001532) and Latino (singleton) individuals. Interestingly, one haplotype has high homology with a singleton haplotype found in an individual in the 1000 Genomes Project from Great Britain. Both the Denisovan and the British singleton O haplotype are on A backgrounds with the exception of a single nucleotide change at position 9:136133506.

The Neanderthal individual is homozygous for two rare "non-deletional" O haplotypes (Figure 2.6). While the non-deletional O haplotype is found at a higher frequency in contemporary European populations the Neanderthal individual was homozygous for the same non-del O haplotype. This Neanderthal non-O haplotype resembles a singleton haplotype found in an individual in the 1,000 Genomes Project from the

Iberian Peninsula. Both the Neanderthal and Iberian Non-del O haplotype share high sequence similarity only differing at 3 loci (9:136,132,873, 9:136,133,506, and 9:136,137,547).

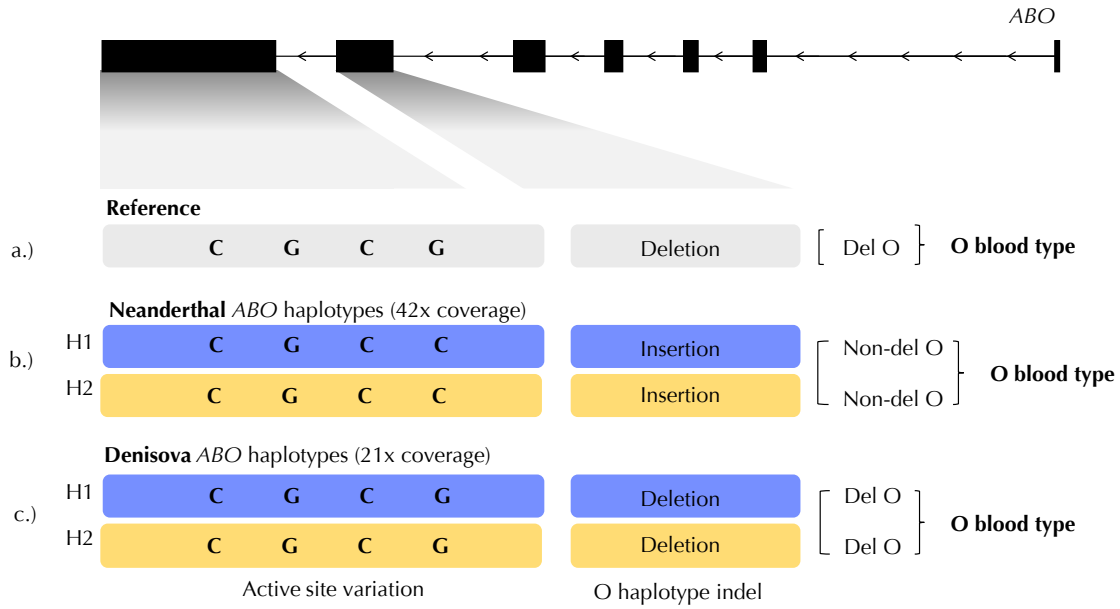

Figure 2.6: **Ancient hominid *ABO* haplotype structure.** Above is an illustration of the *ABO* gene depicted in reverse strand form; ABO exons are depicted as black rectangles. Panels A, B, and C are blowups of *ABO* exons 6 and 7. Panel A shows the reference as a single haplotype. Panel B shows both Neanderthal *ABO* chromosomes. Panel (c) shows both Denisovan *ABO* chromosomes, while (a) shows the reference chromosome depicted as an O haplotype with the O deletion in exon 6 accompanied by the A haplotype background in all four active site loci in exon 7. Panel (b) shows that the Neanderthal individual is homozygous for two rare "non-deletional" O haplotypes as illustrated in both copies of the *ABO* exon 7, active site loci. This Neanderthal individual is homozygous for the same non-del O haplotype not found in any contemporary human population. The Denisovan individual has two unique forms of deletional O, not found in any contemporary human population. Both of these Denisovan O haplotypes are on A backgrounds. Both ancient hominid VCFs were merged and combined with the 1000 Genomes Project VCF (69 loci total) Haplotypes were then called using PHASE 2.1.2. After pairing haplotypes both ancient hominids typed as O blood type.

2.4 Discussion

2.4.1 Imputing ABO blood type for the 1,000 Genomes Project

After constructing ABO haplotypes for the 1,000 Genomes Project dataset we then paired those haplotypes and imputed ABO blood type based on combination assignments of A, B, and O haplotypes (see Figure 2.5 b and Supplemental Table 2.4 for imputed blood type in 5 super populations, and Supplemental Table 2.6 for frequencies and haplotype counts in all 26 populations). Some interesting trends that recapitulate established serologically derived observations from canonical ABO blood group research (e.g., Mordant and Cavalli-Sforza) include: (1) the highest rates of the O blood type are found in admixed American populations, i.e., South America (65%); (2) the absence of the weak A subtype in East Asia (0%); (3) the high frequency of the B blood type in Southern Asia (34.2%); (4) the highest rates of the both A sub types, A and weak A, being found in Europe, 28.2% and 13% respectively; and (5) the high level of haplotype diversity found in Africa (58 haplotypes total, almost twice as many unique haplotypes as any other group included in the 1,000 Genomes Project) (see Table 2.2). Given the demographic history of African haplotype structures resulting in shorter haplotype blocks when compared to younger populations (e.g., Europeans) who have undergone less recombination events, we expect to see the highest levels of genetic variation among the five super-populations included in the 1,000 Genomes Project. This result is consistent with existing population genetic research both on a global scale and in Africa (Tishkoff et al., 2009).

2.4.2 Discovering rare ABO haplotypes

Within our combined exome dataset, we were able to discover three unidentified potential “weak B” haplotypes. These weak B haplotypes are defined by a combination of both the common B haplotype background and the A2 subtype indel (rs56392308, deletion). Within our combined exome dataset eight individuals shared at least one weak

B haplotype. With three of those eight being homozygous for the A2 subtype indel (rs56392308, deletion) and putatively typing as serological blood type AB. The haplotypes defined in this study were statistically derived with corresponding estimations of error and if needed, molecular confirmation of particularly clinically important haplotypes i.e., potential "weak B" will be considered in the future. Finally, upon searching the 40 B subtype alleles on the Blood Group Antigen Gene Mutation database (BGMUT) no alleles submitted had the Weak B haplotype structure described above (Patnaik et al., 2014). Within the 1000 Genomes Project dataset we likely discovered a rare cis-AB haplotype (n=1). These haplotypes are extremely rare and mostly found in East Asian populations (Patnaik et al., 2014). As a result, it is intriguing that a haplotype this rare was identified in an individual of African Caribbean ancestry from Barbados and not in the East Asian cohort included in the 1,000 Genomes Project dataset (see Supplemental Table 2.4 for a detailed look at each haplotype).

2.4.3 Singleton variation segregates on LOF haplotype backgrounds

Of the 8 singletons identified in our combined exome sequencing dataset 7 (87.5%) segregate on O haplotype backgrounds (see Table 2.3). Interestingly the singleton identified in an individual of European Ancestry on a B background within our combined exome dataset (chr9:136132867) is also present on the exome aggregation consortium database (ExAC) as a singleton found in another European individual but does not have a dbSNP entry.

Of the 20 singletons identified in in the 1,000 Genome Project Dataset, 18 or (90%) were also identified on O haplotype backgrounds. The remaining 2 haplotypes or (10%) were identified on classic A haplotype backgrounds. Both of the individuals identified with A singleton haplotypes were of Han Chinese ancestry. One of those Han Chinese specific singleton alleles (rs551612515) was present on ExAC and identified in 5 East Asian individuals. However, the remaining Han Chinese specific A singleton haplotype allele was not present on ExAC or dbSNP (chr9:136131556) and represents an interesting

variant for future validation experiments as it is located in exon 2 and might result in a novel LOF haplotype.

One interesting observation in both datasets is that the majority of both rare and singleton variation occurs on O haplotype backgrounds possibly because these singleton variants accumulated after these O haplotypes had already lost function and thus this rare variation would not have an effect on ABO phenotype. It is also possible that when both singletons and rare variation have a phenotypic effect it might result in LOF itself (e.g., non-del O haplotypes). One explanation might be that because no protein is being expressed there is no additional possibly deleterious phenotype resulting from these variants for the population (MacArthur et al., 2012). However, this is conjecture that requires follow up experiments in the form of analysis of larger datasets accompanied by molecular confirmation of rare variants with unknown significance.

2.4.4 Ancient hominid *ABO* haplotype structure

To assess whether the Neanderthal non-deletional O haplotype is present in more ancient primate lineages we searched three non-human primate coding sequences for non-deletional O haplotypes (Ségurel et al., 2012). We were not able to identify the non-deletional O haplotype in the three primate lineages (chimpanzees, gorillas, and gibbons), possibly suggesting a more recent accumulation of a non-deletional loss of function on the A haplotype background; however, a more exhaustive analysis is needed to confirm this hypothesis as only a limited survey of non-human primate genetic variation exists (Sudmant et al., 2013). My results indicate the existence of a homozygous non-deletional O haplotype in the hominid lineage that is present at low frequencies in modern human populations a finding which is consistent with previous estimates of consanguinity within the Neanderthal lineage (Prüfer et al., 2014). This Neanderthal specific non-del O haplotype also shares sequence similarity with an Iberian individual found in the 1,000 Genomes Project dataset and it is well known that the non-del O haplotype is found at its highest frequency in contemporary European populations. It is

difficult to speculate on the potential for Neanderthal specific *ABO* haplotypes as we have a limited sample size. However the trend is interesting given that Europeans share more Neanderthal sequence than any other population and previous research from Akey et al. supports *Homo sapiens* and Neanderthal population introgression (Vernot et al., 2016).

Another interesting observation form this ancient hominid analysis is that much like the majority of the samples included in this study (with the exception of one chromosome identified in the 1,000 Genomes project that segregates on a B haplotype background) both of the Denisovan O haplotypes identified occur on A haplotype backgrounds. This provides additional evidence supporting that the A haplotype occurred first in the primate lineage with the O frameshift occurring later on the A haplotype background resulting in the loss of function we know today (Saitou and Yamamoto, 1997). However, aDNA has plenty of limitations including short fragment sizes, chemical transformations, and finally our limited sample size (n=2).

2.4.5 The origins of *ABO* haplotypes

Since the discovery of the ABO blood group system over 30 additional blood group systems have been identified (Patnaik et al., 2014). The scale and assortment of variation is highly variable depending on the antigen-presenting gene but according to the NCBI's blood group gene mutation database (BGMUT), varies from 2 to greater than 100 known alleles per blood group system with the majority of common variation found in the form of missense and nonsense SNVs (Patnaik et al., 2014). Allelic diversity in some of these genes has been associated with susceptibility to infectious disease. Well-established links such as, allelic variation in *ABO* have been shown to alter susceptibility to malaria for example (J. M. Moulds and Moulds, 2000).

Such findings support the hypothesis where different blood group antigen genes serve as "incidental receptors for viruses and bacteria" (J. M. Moulds, Nowicki, Moulds,

and Nowicki, 1996), but also function as “modulators of innate immune response” (Lindén et al., 2008) and possibly as “decoy-sink” molecules targeting pathogens to macrophages (Gagneux and Varki, 1999). Given these established observations and the phenomenon where host–pathogen interactions are major determinants of natural selection, antigen presenting genes like *ABO* are often thought of as possible targets of diverse selective pressure (Saitou and Yamamoto, 1997). This observation is in agreement with the geographic differences we observe in allelic diversity in antigen presenting genes, i.e., difference in allele frequency with canonical research supporting non-neutral modes of natural selection namely balancing selection in genes such as the *ABO* loci (Saitou and Yamamoto, 1997). In summary the genotypic diversity and phenotypic variation observed in antigen presence such as A, B, and O among various human populations is believed to be the result of pathogen-driven selection (Saitou and Yamamoto, 1997). As a result human population frequencies of *ABO* haplotypes vary globally thanks to balancing selection (Charlesworth, 2006).

2.4.6 *ABO* balancing selection and trans-species polymorphisms

In addition to its clinical utility our exploration of rare variation and haplotype structure in *ABO* may also create a more comprehensive understanding of balancing selection in the human genome, which is believed to be the selective pressure that shaped *ABO* diversity throughout primate history (Georges et al., 2012). The history of variation in the ABO gene is ripe with controversy concerning which haplotypes (A, B, or O) arose first and how many recombination events have occurred in the *ABO* gene over the course of human history (Yamamoto et al., 2014).

This brings us to a central question in *ABO* population genetics; when did the evolutionary conservation of “balanced alleles” in *ABO* occur? Is it specific to the human lineage? If so which haplotype originated first: A? B? or O? Was this “reappearance” of A and B antigens was first? Was it the result of an ancient polymorphism preserved across species or both? Or was it due to multiple, more recent instances of convergent

evolution? Answers to these questions have been hotly debated for decades (Yamamoto et al., 2012). Until very recently the consensus was in support of "convergent evolution" as the driving force behind the variation found in *ABO* (Leffler et al., 2013). Recent contributions from Molly Prezorski's group support the hypothesis that the effects of balancing selection resulting in the conservation of critical amino acid substitutions (i.e., A, B, and O) can be found in the form of trans-species polymorphisms in many non-human primates, i.e., variation in *ABO* is conserved and not restricted to human primates (Ségurel et al., 2012).

Human genetic variation in *ABO* has been described previously (Cavalli-Sforza et al., 1964); however, here I provide an in-depth description of the distribution of genetic variation based on NGS-derived *ABO* genotypes. My analysis resolves *ABO* haplotypes and imputes A, B, AB, and O phenotypes from almost 9,000 contemporary humans and two archaic hominids. I have shown that it is possible to mine variation in the coding portion of the *ABO* gene from multiple NGS platforms and accurately characterize both common and rare haplotype diversity. While these results are promising and recapitulate previous estimates of *ABO* haplotypes in global populations, at present, we still require correlating serological data to better understand the functional effects of this genetic variation. To this end, in the next chapter I will compare NGS-based ABO haplotypes to serologically determined ABO blood types and describe the potential for these efforts in a clinical setting.

(THIS PAGE LEFT INTENTIONALLY BLANK)

Chapter 3

BUILDING A SEQUENCE-BASED CALLER FOR *ABO*

Analysis of *ABO* haplotypes derived from NGS data holds promise for the development of higher-resolution ABO blood-typing diagnostics. Previous studies attempting to determine ABO blood type from genetic variation in *ABO* have genotyped specific *ABO* SNVs rather than defining the haplotype sequence necessary for subtype detection. I have developed an accurate ranking method to assign *ABO* haplotype (i.e., subtype) using DNA sequence data. This new approach (ABO-Seq) uses variation prioritization to compare phased sequence data in the coding portion of *ABO* to reference A, B, and O haplotypes from the Blood Gene Mutation Database (BGMUT). I apply this to phased NGS data from multiple sequence datasets derived from two distinct sequencing platforms: (1) whole exome (the NHLBI-ESP and MH-GRID, n=6,432) and (2) custom/targeted capture (the Bloodworks Northwest Blood-Seq Project, n=1,140). I developed and tested the algorithm in those NHLBI-ESP participants with both DNA sequence and ABO serologic phenotype data (n=80). I then trained the algorithm on 467 individuals within the Bloodworks Northwest's Blood-Seq dataset achieving ~99% concordance with ABO serologic phenotype. The five discordant samples were then investigated for the presence of other genetic variation in *ABO* and used to refine the ABO-Seq algorithm to detect (1) other non-deletional O haplotypes and (2) rare variation. We identified common *ABO* variants known to influence function, including the common exon 6 indel that identifies *ABO O-01* genotypes, and another common exon 7 indel that results in common *ABO*A2* genotypes and and A2 serology subtype. We have also identified rare coding variants within *ABO* (single nucleotide/missense variants, insertion/deletions) that segregate on common haplotype blocks. This approach (ABO-Seq) has the potential to improve the resolution of ABO blood type determination at both the clinical and research level and to reveal novel associations between disease phenotypes and *ABO* genetic variation.

3.1 Introduction

Next-generation sequencing (NGS) is rapidly making in-roads into clinical practice (Johnsen, Nickerson, and Reiner, 2013b), yet it's potential to transform current standard methods such as antibody-based ABO blood typing has yet to be fully realized. The *ABO* gene encodes the glycosyltransferase that adds A or B sugars to the H antigen substrate.

Single nucleotide variation (SNV) in the *ABO* gene affects the function of this glycosyltransferase at the molecular level by altering enzyme specificity for sugar donors and efficiency (Storry and Olsson, 2004). Characterizing genetic variation in *ABO* is of great interest to the transfusion and transplantation medicine community as variant ABO phenotypes can have significant consequences with regard to donor-recipient compatibility (Storry and Olsson, 2004). Relating *ABO* genotypes to actual blood antigen phenotype requires the sequence analysis of haplotypes (Scheet and Stephens, 2006). These haplotypes are composed of both common and rare DNA variants predicted to affect the structure and or function of the *ABO* glycosyltransferase.

Moreover, many large-scale population based screens have immense value as they have not only used NGS methods to sequence diverse communities of people. The A1 subtype (the reference blood type for the ABO blood group) has a robust serological clumping pattern that can sometimes distinguish it from other, weaker A subtypes in part because the density of A antigens on the red blood cell surface differ substantially (Yamamoto, McNeill, and Hakomori, 1992). A common A2 glycosyltransferase is a result of a single nucleotide deletion in the 3' coding portion of the *ABO* gene. This allelic variation can be characterized with higher resolution using sequence based techniques (e.g., *ABO*A1.01*, *ABO*A2.01*, *ABO*A2.06*, etc.) rather than antibody-informed categories of ABO type. According to the Blood-group Antigen Gene Mutation Database (BGMUT), in the *ABO* gene there are at least 300 known allelic variants discovered with various methods (Patnaik, Helmberg, and Blumenfeld, 2014). Many of these *ABO* alleles are reported accompanied by ABO serology and most are nonsynonymous variants thought to affect ABO glycosyltransferase structure and function (Patnaik et al., 2014).

Our goal was to utilize both serologic phenotype data and detailed DNA sequence-based assessment of common and rare variation in the form of *ABO* haplotypes to (1) increase granularity in ABO determination, and (2) develop the basis for future high resolution ABO disease association studies based on variation in the coding portion of

the *ABO* gene, allowing for sequence based imputation/interpretation of ABO subtypes in large-scale NGS datasets.

The methods and analyses discussed in this chapter provide a detailed view of both common and rare variation present in the coding portion of the *ABO* gene that predicts different ABO serologic patterns in 1,220 individuals. Both NGS datasets included in this study are accompanied by serological ABO phenotype data allowing for confidence in the high resolution assignment of *ABO* blood type (subtypes) *via* haplotypes composed of rare and common variation in the coding portion of the *ABO* gene. In this study we (1) detail the development of a method to predict ABO type from DNA sequence data extracted from the coding portion of the *ABO* gene (ABO-Seq), (2) assess the ABO-Seq algorithm by comparing our DNA sequence based results/calls to serology defined ABO blood types, (3) discuss both the success and limitations inherent in sequence based ABO blood type determination and finally (4) discuss the future of adopting sequence based ABO blood typing in practice.

3.2 Methods and subjects

3.2.1 BloodSeq: targeted capture of 41 human blood-group antigen genes

In collaboration with Bloodworks Northwest we developed an NGS approach to generate targeted capture data for 1,140 blood donors, phenotyped previously by serology for 35 different blood-group antigen systems (including ABO serology). NGS sequence data was generated using a targeted sequencing panel and Illumina paired-end 100 bp reads for 41 blood-group genes including exons and introns (Supplemental Table 3.1). Through BloodSeq, sequence data was generated for 2,280 *ABO* chromosomes at high coverage (>100x) in a total of 1,140 consenting blood donors self-identified to be of Asian American descent who donated blood in the Puget Sound Blood Center system spanning the western region of Washington state. The library construction, targeted capture, sequencing, mapping, calling, and filtering were carried out as described previously

(Gordon et al., 2016). It is important to note that these samples were enriched for O blood type in order to increase the probability of detecting a variety of loss of function *ABO* alleles (Personal communication, Jill Johnsen, 2015). Thus, the distribution of A, B, AB, and O serological blood types were selected *a priori* and as a result do not represent the natural population distribution of ABO blood types in these populations. Finally, linking NGS derived genotype to ABO phenotype we have serology data for 467 individuals included in the Bloodseq dataset (i.e., training set 2, Table 3.1)

3.2.2 Aggregating multiple human exome sequencing datasets

We used *ABO* coding sequence data derived from the NHLBI Exome Sequencing Project (ESP) and the Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID). Through the ESP, 15,336 genes were sequenced at high coverage (median depth >100x) in a total of 6,515 unrelated European American (EA, n=4,298) and African American (AA, n=2,217) individuals from 19 different cohorts. Through MH-GRID, exome sequences at similar high coverage (median depth >100x) were obtained from a total of 1,313 unrelated AA individuals. The library construction, exome capture, sequencing, mapping, calling, and filtering were carried out as described previously (Johnsen, Auer, Morrison, Jiao, et al., 2013a). Exome sequence data were aligned to NCBI human reference GRCh37.

Sequence data were aligned to reference sequence obtained from GenBank (NG_006669.1). Average sample read depth for the *ABO* gene was 77x (ranging from 10 to 374x) and includes the entire coding sequence and exon-intron boundaries (see Table 3.1). Of the 4,298 EA ESP participants, 3,405 had minimum coverage of 50x across the targeted *ABO* exons. Additionally, 3,027 self-identified African Americans (AA) from both the ESP and MH-GRID were included in our *ABO* analysis bringing our total to 6,432 participants, roughly equally divided between EA (53%) and AA (47%). Finally, linking NGS derived genotype to ABO phenotype we have serology data for 80 individuals included in the ESP dataset (i.e., Training Set 1, Table 3.1)

Table 3.1: ***ABO* variants identified in both datasets were limited to coding variation derived from VCFs generated using GATK (McKenna et al., 2010).** A sub-set of both datasets included in this study have corresponding ABO serologic phenotype data (i.e., training set1, and training set 2).

Dataset	NGS platform	Depth	Samples	Ethnicity	Missense	Frameshift	Splice	ABO serology
NHLBI-ESP + MH-GRID	exome	~100x	6,432	European American and African American	45	7	2	80
Blood-Seq	targeted capture	~77x	1,140	Asian American	31	3	2	1140

3.2.3 Curating *ABO* alleles on BGMUT

The BGMUT is a National Center for Biotechnology Information (NCBI) administered information repository for blood-group antigen system alleles (Patnaik et al., 2014). However, a major challenge that will hinder the application of sequence-based ABO typing is the translation of allelic variation cataloged in the BGMUT database into genome-based coordinates. Many of the *ABO* alleles in BGMUT were derived by a variety of methods (e.g., RFLP, PCR, Sanger sequence etc.) and as a user-deposited database there is variation in annotation formats. In my attempts to convert BGMUT *ABO* alleles to genomic coordinates, some BGMUT entries did not correspond to the *ABO* allele that was entered in the database. Moreover, the *ABO* alleles specified did not correspond to the amino acid specified in the entry and were not able to be verified by location by querying dbSNP. This reduced the number of alleles in the database which could be reliably translated to genomic coordinates in an automated fashion by half (151/377). In order to refine and convert *ABO* reference alleles we filtered each FASTA that exists on the on BGMUT for size and uniqueness (Patnaik et al., 2014) (see supplemental figures 3.2 a and 3.2 b).

3.2.4 BGMUT and Conversion of cDNA positions to genomic coordinates

Alignment and conversion of *ABO* allele entries from cDNA to genomic coordinates is essential for conversion of BGMUT allelic variation to reference alleles in hg19 coordinates, many of the BGMUT allele entries are rare (MAF <1%), and finally of utmost importance have corresponding serologic ABO phenotype data of value when annotating variants discovered in NGS datasets. Of the 377 *ABO* allele entries on BGMUT, 151 had DNA sequence data meeting criteria for this analysis. *ABO* allele FASTAs downloaded from BGMUT were then aligned to the *ABO* gene reference sequence using crossmatch (la Bastide and McCombie, 2007). We then compared and counted loci that differed when comparing BGMUT allele FASTA to the *ABO* reference, yielding BGMUT allele variant files in hg19 coordinate lookup tables. These lookup tables were then compiled into a multi-sample variant calling format (VCF) file that includes 151 BGMUT allele entries. 90 total SNVs were identified in the coding portion of *ABO* through this analysis (see Supplemental Table 3.2 for annotated BGMUT variant loci in hg19 coordinates, n=151 BGMUT lookup tables).

3.2.5 Resolving *ABO* haplotypes in NGS datasets and imputing ABO blood type

In order to resolve *ABO* haplotypes from NGS data, we employ PHASE 2.1.2 for haplotype construction of the different chromosomal alleles for each individual (Scheet and Stephens, 2006). We then compare these phased *ABO* haplotypes to A, B, or O reference haplotypes defined by FASTA files curated for uniqueness downloaded from the NCBI BGMUT reference repository (Patnaik et al., 2014). We then use an in house matrix scoring method where each collective reference and variant base genotype across our BGMUT reference lookup table is compared to the sequence derived phased genotype as shown in Figure 3.1.

We first resolve critical haplotypes using six loci known to be responsible for common amino acid substitutions (i.e., rs7853989 (p. R176G), rs8176743 (p. G235S),

rs8176746 (p. L266M), and rs8176747 (p. G286A)) and two frameshift indels one responsible for the common loss of function *ABO*O.01* haplotypes in exon 6 (*ABO* c.261delG) and the second an indel at the end terminus of exon 7 (*ABO* c.1061delC) responsible for the addition of +21 amino acids and the A2 weak hemagglutinin phenotype. After comparison across the matrix of reference haplotypes for critical variation of 6 essential loci (first LOF, then Missense variation in the active site, then potential weak serology alleles) these scores are ranked and the highest match count haplotype (first LOF, then Missense variation in the active site, then potential weak serology alleles)is chosen as that phased allele for the individual as shown in Figure 3.2. Finally, we then pair those predicted A, B, and O haplotypes and predict ABO blood type.

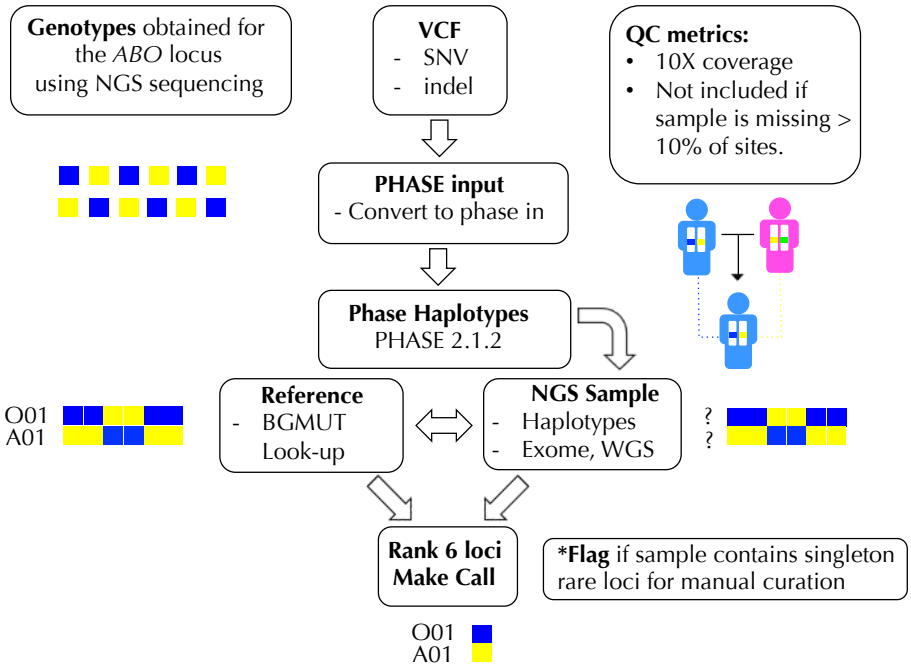

Figure 3.1 **Overview of ABO-Seq pipeline.** We first obtain genotypes in the coding portion of the *ABO* gene in the form of a multi-sample VCF file. Quality control metrics are essential for VCF accuracy. If coverage is less than 10x, variant calling quality is poor for any of the 6 positions used to define haplotype classes, or >10% of the *ABO* data has poor variant calling quality the sample is not included. We then use a custom PHASE 2.1.2 input script to convert our multi-sample VCF to

PHASE input. We then run PHASE to resolve *ABO* haplotypes. We next compare our newly phased *ABO* sample haplotypes to our BGMUT lookup table for all 6 loci, i.e., the O (c.261delG), the A2 (c.1061delC), and the 4 active site missense variants and count and score similar loci. Finally, we rank those 6 loci prioritizing the presence of c.261delG and make an ABO prediction. It is important to note that if a phased ABO haplotype is a singleton, or has rare variation on a common *ABO* haplotype background it is investigated further via manual curation.

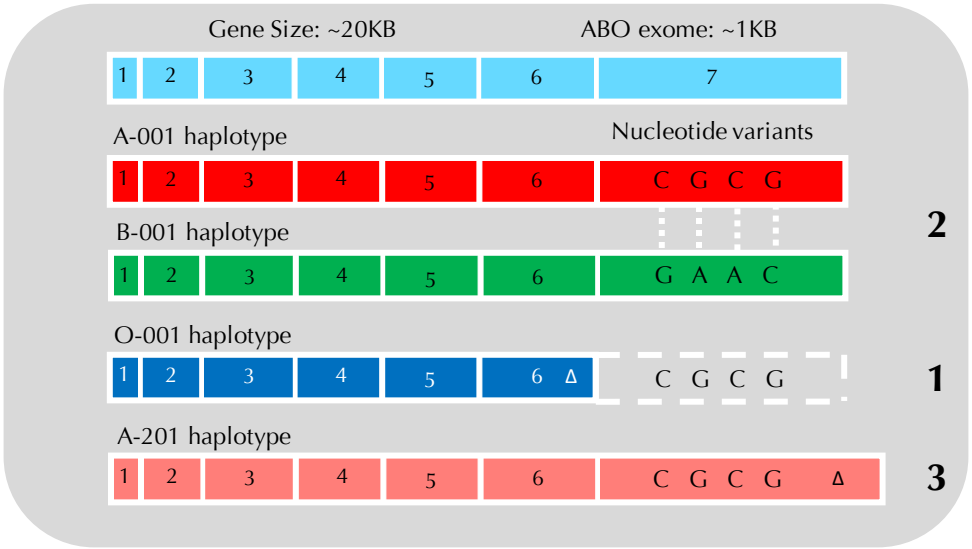

Figure 3.2 **Overview of *ABO* haplotype comparison and ranking scheme.** We first resolve critical *ABO* haplotypes using six loci known to be responsible for the common amino acid substitutions (i.e., rs7853989 (p. R176G), rs8176743 (p. G235S), rs8176746 (p. L266M), and rs8176747 (p. G286A)) and two frameshift indels one responsible for the common loss of function O1 haplotype in exon 6 and the second an indel at the end terminus of exon 7 responsible for the addition of +21 amino acids and an A2 weak hemagglutinin phenotype. (1) trumping mechanism; If the exon 6 O LOF indel is present it trumps all variation occurring after the LOF indel and results in an O haplotype (2) Ranking Active site loci; If one of these 4 active site loci in our haplotype resolved NGS data matches loci in the BGMUT allele lookup table that reference haplotypes receives a +1 score. (3) After all sites are compared across the matrix of reference haplotypes for critical variation of those 6 essential loci these scores are ranked and the highest match count haplotype is chosen as that phased allele for the individual as shown above. Next we assess the weak serology indel loci. Finally, we then pair those predicted A, B, and O haplotypes and impute ABO blood type.

3.2.6 **Rare variation and singleton analysis**

After assessing the variation in those 6 loci, we build our final *ABO* haplotypes using all coding variation present in a given dataset including rare and singleton variation. This allows us to assess upon which common haplotype background rare variation occurs. We have assembled haplotypes for both NGS datasets included in this study ranging from 85 variant coding loci (combined exome dataset) to 53 loci (Blood-Seq) in those same regions. In the case of rare/singleton variation processed samples are tagged and manually compared to variation identified in to BGMUT allele reference lookup tables (see Supplemental Table 2).

If those loci are not present in our BGMUT lookup table database (i.e., BGMUT multi-sample VCF) we searched for those loci in the exome aggregation consortium database (ExAC) (Minikel et al., 2016), to evaluate (1) allele frequency and human population diversity and (2) location of the rare/singleton variant on the *ABO* gene to evaluate its potential effect on structure or function using knowledge of *ABO* variation related to ABO protein domain functions (Yamamoto et al., 2014) and variation annotation heuristics such as CADD (multiple heuristics), GERP (conservation), and POLYPHEN-2 (amino acid impact) (Adzhubei, Jordan, and Sunyaev, 2013; Cooper et al., 2005; Kircher et al., 2014). For example, if a rare/singleton variant is not present in BGMUT, ExAC, or dbSNP it will be further evaluated for potential LOF (CADD, GERP, etc.) and location relative to ABO functional domains. If the sample harboring the rare variant has corresponding informative ABO serology then additional curation of the variant annotation and refinement of prediction algorithms can be performed.

Finally, similar approaches to predict ABO blood type from NGS data exist and have attempted to utilize boolean methods (i.e., strict counting) to determine ABO blood type with little accuracy because their methods do not prioritize or evaluate variation present in the *ABO* gene that effect structure/function on the biological level (Giollo et al., 2015). Our method prioritizes variation in the coding portion of the *ABO* gene that might alter

the expression or structure/function of the ABO glycosyltransferase. Moreover, our method also allows for the construction of *ABO* haplotypes composed of as many as 85 variants resulting in a method to assign high resolution ABO alleles and corresponding subtypes from haplotypes that include rare and singleton variation.

3.3 Results

In our combined exome dataset consisting of samples from both the NHLBI-ESP and the MH-GRID, we investigated *ABO* variation in 6,432 diploid participants (i.e., the coding sequence of 12,864 *ABO* chromosomes). The data were stratified genetically using a principal components analysis. Among the exome samples, 3,405 samples are of European ancestry (EA) and 3,027 are of African ancestry (AA). Of the 74 non-synonymous variants identified; 16 were common (MAF <10%), 6 had an MAF between 1-10%, and 53 were infrequent, or rare, with an MAF >1%. 12 of these rare variants were singletons, 4 were doubletons, and the remaining 37 had an MAF of <1%. A subset of these *ABO* variants potentially affect the structure and/or function of *ABO* glycosyltransferase. These non-synonymous variants categorically include missense variants resulting in potential amino acid substitutions (n=45), nonsense/frame shifts (n=7), and splice site variation (n=2).

In our Blood-Seq targeted capture dataset, we investigated *ABO* variation in 1,140 participants (i.e., the coding sequence of 2,280 *ABO* genes). All of the participants included in this dataset were of self-identified Asian American ancestry. Of the 55 SNVs identified in the coding portion of the *ABO* gene; 16 were common (MAF >10%), 5 had an MAF between 1-10%, and 33 were infrequent, or rare, with an MAF <1%. 18 of these rare variants were singletons, 8 were doubletons, and the remaining 37 had an MAF of <1%. Non-synonymous variants include missense variants (n=31), nonsense/frame shifts (n=3), and splice site variation (n=2). The full spectrum of human genetic variation

identified in the coding portion of *ABO* in all 1,140 individuals included in the Blood-Seq dataset is summarized on a per exon basis in Figure 3.5, Supplemental Table 3.1.

3.3.1 SNVs in 151 BGMUT alleles

Within our BGMUT multi-sample VCF composed of 151 ABO allele entries we identified 90 SNVs in the coding portion of the *ABO* gene. We used the ExAC database to query allele frequencies for all 90 ABO variants. 13 were common (MAF >101%), and 74 were infrequent, or rare, with an MAF <1%. 13 of these rare variants were singletons, 3 were doubletons, and 40 were not reported in ExAC (Minikel et al., 2016). Non-synonymous variants categorically include missense variants resulting in amino acid substitutions (n=69) and nonsense/frame shifts (n=7). The full spectrum of human genetic variation identified in the coding portion of *ABO* in all 151 *ABO* allele entries included in the BGMUT database is compared to the Blood-seq dataset and summarized on a per exon basis in Figure 3.4, Supplemental Table 3.2.

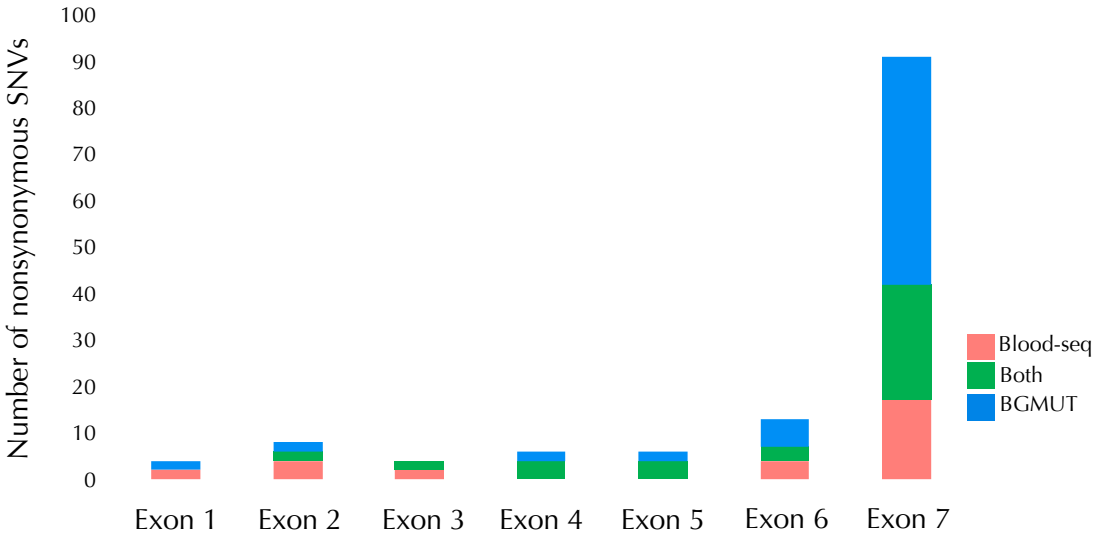

Figure 3.3: **Comparing *ABO* SNVs in the BloodSeq to the BGMUT database by exon.** Of the 90 *ABO* coding SNVs identified in 151 eligible *ABO* entries in the BGMUT database, 32 of those nonsynonymous loci were included in both datasets, 58 SNVs were found exclusively in the BGMUT dataset and 23 SNVs found exclusively in the Blood-Seq targeted capture dataset (for a comprehensive list of variants in both datasets as well as hg19 coordinates see Supplemental Table 3.3). The majority of variation occurs in exon 7 of the *ABO* gene. This is consistent with previous studies on the accumulation of nonsynonymous variation in the final exon of most genes in the human genome (MacArthur et al., 2012).

3.3.2 High-throughput exploration of *ABO* haplotype diversity

ABO haplotypes are composed of both common and rare variation in both datasets. Each haplotype in our combined exome sequence dataset is informed by 85 variable loci. In our combined exome dataset of 6,432 individuals we characterized 211 unique *ABO* haplotypes from 12,864 *ABO* genes (Table 3.2). 114 of the 211 (54%) haplotypes discovered were singletons. 23 of the 211 were doubletons r4i8(11%), and 16 of the 211 were tripletons (8%). Of the 211 unique *ABO* haplotypes, 135 (64%) included the O c.261delG (rs8176719), 54 (26%) included an A predicted haplotype background, and 22 (10%) included B haplotype defining variants. Of the 12,864 *ABO* genes exome

sequenced, 8,610 were classified as O haplotypes, 2,934 were classified as A haplotypes, and 1,320 were classified as B haplotypes. Within the common A haplotype group, 19 of the 54 or 34% included the A2 subtype indel c.1061delC (rs56392308, deletion).

3.3.3 Haplotype structure of 1,140 Asian individuals in the Blood-Seq dataset

In our BloodSeq targeted capture dataset we characterized 62 unique *ABO* haplotypes from 2,280 *ABO* chromosomes (Table 3.2, Supplemental Table 3.4). Each haplotype is composed of 53 variable loci. 31 of the 62 (50%) haplotypes discovered were singletons, 15 of the 62 were doubletons (24.1%), and 1 of the 62 were tripletons (1.6%). Of the 62 unique *ABO* haplotypes 38, or 61.2% included the O c.261delG (rs8176719, deletion). When characterizing common haplotype backgrounds, 14 or 22.5% included common A haplotype backgrounds, and 5 or 8% included common B haplotype backgrounds. Of the 2,280, *ABO* genes sequenced 1,462 were classified as O haplotypes, 484 were classified as A haplotypes, and 316 were classified as B haplotypes. Within the A haplotype group, 2 of the 14 or 14.2% included the A2 c.1061delC (rs56392308). We also identified 4 rare non-deletional O haplotypes (15 chromosomes total) and 1 possible recombinant loss of function *ABO* haplotype.

Table 3.2: **Summary of *ABO* haplotypes in both our combined exome dataset and the Blood-seq targeted capture dataset**. Summary of *ABO* haplotypes constructed using PHASE 2.1.2. Each cohort includes unique haplotypes that are then further broken down into A1, non-A1 A, B, O01, non-deletional O, and Cis-AB haplotypes. In parenthesis we have provided the number of chromosomes identified in each haplotype category.

Cohort	Unique *ABO* haplotype	A1 haplotype	Non-A1 A haplotype	B haplotype	O01 haplotype	Non-del O	Cis-AB
ESP + MH-GRID	211	29 (1,958)	19 (821)	22 (1,320)	135 (8,610)	6 (155)	0
Blood-seq	62	12 (440)	2 (44)	5 (316)	38 (1,462)	4 (16)	2

3.3.4 ABO-Seq/Training set 1: Concordance with typical serological phenotypes

To determine ABO-Seq's accuracy, concordance of exome sequence derived *ABO* haplotypes were compared to (1) serological-derived ABO blood types (n=80, Figure 3.4). Within our first training set, genotype serology ABO-Seq achieved 100% concordance in 80 samples. This early version of the algorithm did not assess rare variation and rather focused on training with the 6 critical variant loci. While we were able to predict ABO blood type with 100% concordance in a small sample size this matrix based ranking approach (i.e., prioritizing common/known amino acid substitutions) has inherent limitations that were highlighted in our second training serology training set of 467 samples selected blinded to ABO serologic type in the Blood-seq dataset.

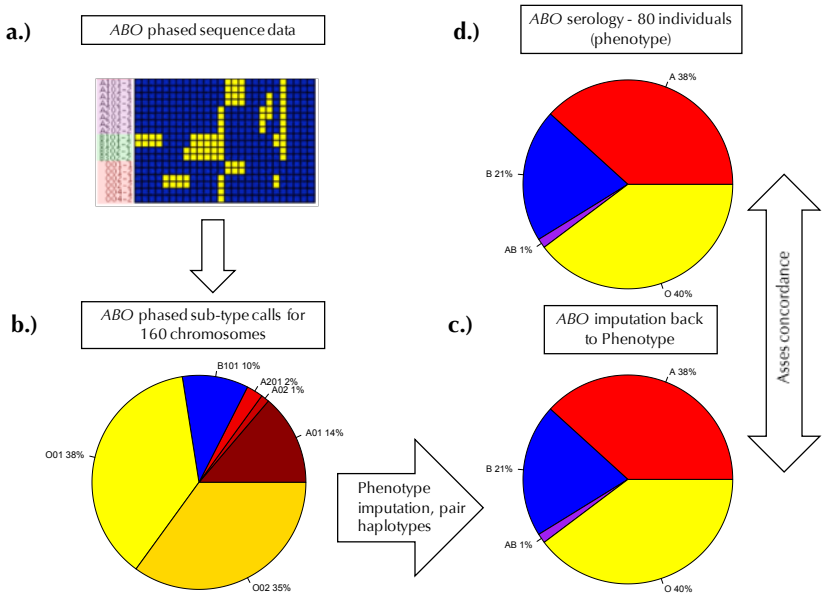

Figure 3.4: **Training set 1 — *ABO* genotyping algorithm with ABO serologic blood type calls.** Phased ABO coding data was used to determine ABO subtype calls for 80 individuals included in the NHLBI-ESP dataset that had corresponding serology (160 chromosomes total). Chromosomes were then paired for each individual (n=80) and putative A, B, AB and O phenotype calls were made. 100% concordance was achieved between ABO-Seq, sequence based calls and serology determined ABO blood type.

3.3.5 ABO-Seq/training set 2 — concordance with ABO phenotypes

Our second phenotype training set (training set 2) included 467 individuals included in the Blood-seq targeted capture dataset. Of the 467 samples, we imputed *ABO* subtype and compared our sequence-based ABO subtype results to ABO serology results (see Table 3.3 and Supplemental Table 3.5). Ninety-nine percent concordance was achieved between ABO-Seq, i.e., sequence-based calls and serology determined ABO blood type within training set 2 (n=467). ABO-seq provided high-resolution *ABO* haplotype/subtype level calls (e.g., Non-Del O, non-A1 A, and Cis-AB subtypes) based on resolving haplotypes and paring them to predict ABO blood type and A subtype class. Moreover, ABO-Seq was able to identify one individual with a rare Asian Specific Cis-AB haplotype (Roubinet, Janvier, and Blancher, 2002). The Cis-AB allele confers a dominant AB phenotype, which although rare is critical to include in algorithms predicting ABO from DNA data.

After resolving and paring *ABO* haplotypes we then imputed ABO subtype and were left with five discordant samples or (1% of samples included in training set 2) (i.e. the serology did not match our prediction in 5 samples).

Table 3.3: **Training set 2 — Concordance of *ABO* haplotypes with *ABO* serology.** ABO-seq calls are compared to serologic determination of ABO blood type in 467 samples. Grey rows indicate A, B, AB, and O predicted phenotypes while white rows indicate *ABO* haplotype resolved subtype class. There were five phenotype-genotype discordant samples (1% of samples included in training set 2).

	ABO-seq	ABO Serology
A	**155**	**154**
A1	144	n/a
Non-A1 A	11	n/a
B	**100**	**99**
AB	**26**	**26**
AB-01	25	n/a
Cis-AB	1	n/a
O	**186**	**188**
O01	183	n/a
Non-del O	3	n/a

3.3.6 ABO Discordant Samples in the Blood-Seq targeted capture dataset

Of the 467 samples included in training set 2, five were discordant between predicted ABO type and serology (see Table 3.4 and Supplemental Table 3.4 for a complete list). Of our five discordant samples ABO-Seq determined, one potentially has a novel loss-of-function active site variant (sample 109001). In an attempt to understand if rare variation was additionally confounding the predictions of the ABO-seq caller, we output a list of variants that were present only in discordant samples. Those variant loci were then analyzed using SeattleSeq annotation. One of those loci was identified as singleton missense variant (9:136,131,247, C>T) in the ExAC database resulting in ABO p.D295N (Figure 3.5). It's GERP score (4.38) and PolyPhen2 score (1.0) indicates that it is conserved and potentially deleterious. Moreover, the individual is O, and therefore this

rare variant must be on an O haplotype highlighting this singleton as a potential rare unidentified non-deletional O haplotype.

Table 3.4: **Discordant samples in training set 2**. Five total discordant samples in the Blood-seq dataset (Training Set 2).

BloodSeq ID	Haplotype 1	Haplotype 2	ABO Allele 1	ABO Allele 2	ABOSeq Call	ABO Serology
107377	25	51	O	B	B	A
113320	25	25	O	O	O	A
107054	1	30	O	A	A	O
109001	25	48	O	A	A	O
113321	11	62	O	Non-A1 A	Non-A1 A	O

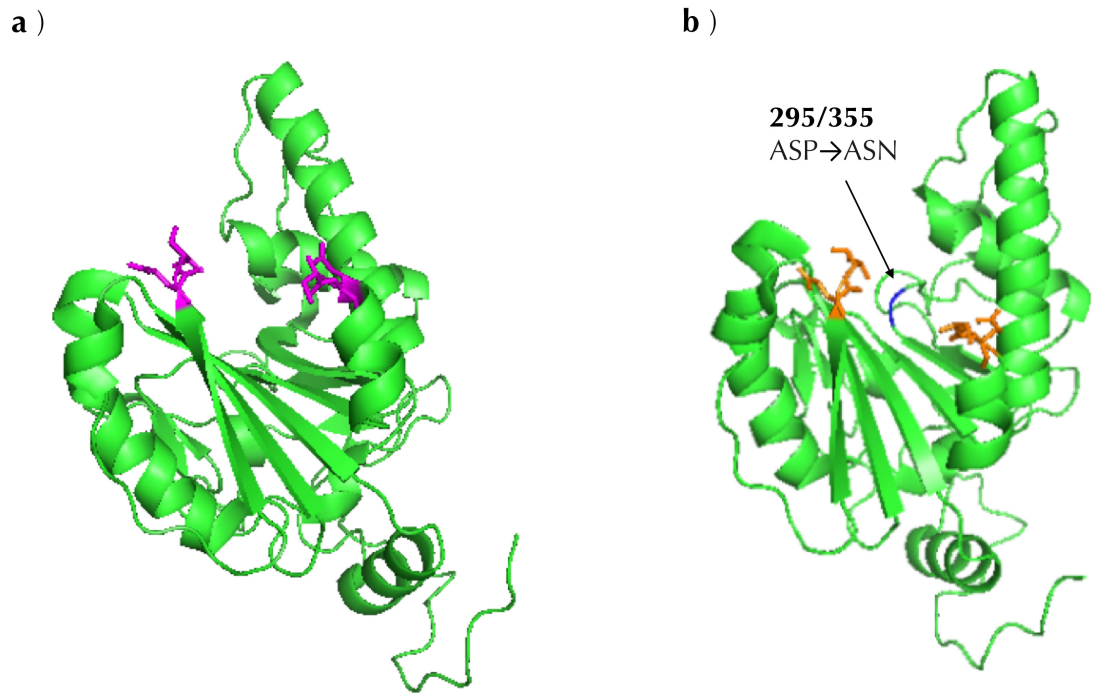

Figure 3.5: **Pymol protein structures of the A isoform of the ABO glycosyltransferase and location of p.Asp295 relative to other critical residues of the ABO glycan binding pocket.** Using the PDB entries for the "A" glycosyltransferase (PDB1lzi) we have retrieved a model of the A glycosyltransferase informed by crystal structure. In panel (A) we have highlighted active site residues which have been shown to differ between A and B isoforms of the ABO glycosyltransferase in magenta and in panel (B) a model of the A glycosyltransferase with the active site loci highlighted in orange and the location of the singleton p.Asp295Asn residue substitution (blue) that was identified in an O01 heterozygous O individual and which is a candidate to be a novel non-deletional O variant identified in training set 2 (sample 109001).

Upon searching the exome aggregation consortium (ExAC), an NGS dataset composed of ~66,000 human exomes from six distinct populations, the *ABO* gene has over 400 coding variant entries, the vast majority of them are rare and population specific (Supplemental Figure 3.3). Upon searching ExAC we identified the same singleton variant p.Asp295Asn in our discordant sample as a singleton found in an individual of Southern Asian descent. Although we do not know if this singleton is responsible for the O phenotype in our discordant sample in training set 2, this variant is a strong candidate

to be a rare Asian-specific *ABO* LOF variant and supports the value of discovery of new variants to better inform ABO prediction from genetic data. This variant has been added to the *ABO* variant lookup table for the next version of the ABO-Seq caller.

We also suspect that ABO-Seq was able to detect a clerical error. One sample 113320 serologically typed as A but ABO-Seq typed it as being homozygous for the O haplotype. Another sample 11321 serologically typed as O but ABO-Seq subtyped it as having one traditional O haplotype, and one non-A1 A haplotype (phasing with 42 other non-A1 A chromosomes). These samples (113320, and 113321) were likely processed one after another and the lack of rare variation as well as their corresponding reciprocal serologic phenotypes indicate that there might have been a sample swap.

The two remaining discordant samples 107054 and 107377. Sample 107377 is serotyped as A, yet has one *ABO* haplotype B101 and one O01 haplotype. Sample 107054 serotyped as O, yet has one *ABO* haplotype is A101 and one O01 haplotype. Upon investigation for known variants in the promoter region (Cai et al., 2013), candidates in intronic sequence, and inspection of the data in IGV for these samples, no candidate variation was identified predicted to result in the loss of function O blood type observed in 107054.

3.3.7 Discussion

The development of our sequence-based ABO caller has led to a better understanding of rare variation in the coding portion of the *ABO* gene. The methods and analysis discussed in this chapter serve as a foundation for future work in search of population-specific rare variation in the coding portion of the *ABO* gene. Our haplotype construction method of assigning and pairing both common and rare *ABO* haplotypes trained in two well curated an unique datasets and could serve as a DNA sequence-based method to assign higher-resolution *ABO* blood types *via* haplotypes. By constructing *ABO* haplotypes from almost 20,000 chromosomes (Chapters 2 and 3), we have demonstrated the

discovery and characterization of both common and novel *ABO* haplotypes in almost 30 populations with potential significance for transfusion and organ transplantation (Gloor et al., 2003). Additionally, a portion of the missing heritability in CVD and cancer-disease-association studies may be due to the low resolution of genetic information used in GWAS disease-association studies which rely upon a limited set of common variants (see Table 3.2) and intronic SNVs reported to be in linkage disequilibrium (LD) with common variation in *ABO* (Wessel et al., 2015). Moreover, 96% of GWAS studies have focused on individuals of European ancestry which have informed definitions of common population-specific variations (Bustamante, Burchard, and la Vega, 2011).

One of the first steps toward the identification of human blood-group genes *via* human genome sequencing will be transitioning from ISBT nomenclature to useful annotation which includes human genome coordinates (hg19 etc.). Clearly, noncoding variants will impact *ABO* gene function. Therefore, our analysis is not a completely comprehensive analysis of all potential functional variation at the *ABO* gene. However, previous research suggests that the vast majority of phenotypic variation in the *ABO* gene is the result of variation in the coding sequence/ While our preliminary results are promising there are still opportunities to continue to discover relevant variation and improve sequence-based *ABO* blood type determination.

For example, the phase state of rare variants is difficult to resolve unless it is on the same read/readpair as a known variant. We are open to exploring additional computational haplotyping methods to optimize the ABO-Seqs's performance in the future (Snyder, Adey, Kitzman, and Shendure, 2015), which can be further improved with curation with phenotypic data and molecular confirmation of variants of unknown significance, novel haplotypes, and emerging sequencing technologies such as long-read platforms (i.e. Pacific Biosystems and Oxford Nanopore).

Finally, ABO provides an opportunty to learn to predict the functional impact of rare variation. We are just starting to realize the effects of rare variation in the human

genome. While we were capable of identifying some common non-deletional O haplotypes, unknown rare non-deletional LOF O haplotypes still present a challenge for future sequence-based determination of ABO blood type (Yazer, Hosseini-Maaf, and Olsson, 2008). Additionally, future iterations of an *ABO* sequence-based caller should include both structural variation that might result in either change of function or loss of function *via* large-scale duplications and deletions (Fox et al., 2016) as well as variation in the *FUT1* and *FUT2* genes where loss of function has been shown to modify ABO by altering H antigen (i.e., the Bombay Phenotype) (Storry et al., 2006). These limitations highlight the importance of current efforts (Bloodwork's Northwest Blood-seq project) to catalog variation in *ABO* by NGS sequencing and the importance of providing corresponding serological phenotype data.

(THIS PAGE LEFT INTENTIONALLY BLANK)

Chapter 4[1]

ANALYSIS OF EXOME-SEQUENCING DATASETS REVEALS STRUCTURAL VARIATION IN THE CODING REGION OF *ABO* IN INDIVIDUALS OF AFRICAN ANCESTRY

ABO[2] is a blood group system of high clinical significance due to the prevalence of ABO variation which in the setting of allogeneic exposure can cause major, potentially life-threatening, complications. Using multiple large-scale next-generation sequence (NGS) datasets, we demonstrate the application of a read-depth approach to discover previously unsuspected structural variation (SV) in the *ABO* gene in individuals of African ancestry. Our analysis of SV in the *ABO* gene across 6,432 exomes reveals a partial deletion in the *ABO* gene in 32 individuals of African Ancestry that predicts a novel O allele. Our study demonstrates the power of large-scale sequencing data, particularly datasets of underrepresented minority populations, to reveal novel SV in the gene responsible for one of the first identified heritable human traits, *ABO*.

4.1 Introduction

The *ABO* gene commonly encodes two different forms of a glycosyltransferase defined by the addition of either the A or B sugars (N-acetyl-D-galactosamine (GalNAc) for A or α-D-galactose (Gal) for B) to the H antigen substrate (L-fusoce (Fuc) (Yamamoto, 1990). Single nucleotide variants in the *ABO* gene affect the efficiency and specificity of this enzyme for these sugars (Yamamoto, 1990). Loss of function (LOF) variation in the form of a single basepair frameshift accounts for the majority of variation resulting in the O phenotype (Petenaude et al., 2002). Characterizing variation in *ABO* is important in transfusion and transplantation medicine because variants in *ABO* have significant consequences with regard to recipient compatibility. Additionally, variation in the *ABO* gene has been associated with cardiovascular disease risk (e.g.,

[1] Portions of this chapter have been adapted, with minor changes, from Fox et al. (2016).

[2] Abbreviations used in this Chapter are as follows: ABO = ABO blood group, SV = structural variation, CNV = copy number variation, LOF = loss of function, RD = read-depth, NGS = next-generation sequencing, SD = segmental duplication, 1KG = 1,000 genomes project, AA = African ancestry

myocardial infarction) and quantitative blood traits (e.g., von Willebrand factor (VWF), Factor VIII (FVIII), Intercellular Adhesion molecule 1 (ICAM-1), E-cadherin, and P-selectin) (Zhang, 2012). Relating *ABO* genotypes to blood group antigen phenotypes requires the analysis of haplotypes. These haplotypes are composed of both common and rare variants that can affect the structure and/or function of the ABO glycosyltransferase (Yip, 2002). Located on the distal end of chromosome 9 (9q34.2), ABO represents an interesting candidate gene for structural variation (SV) discovery as this region has a higher than average recombination rate relative to more proximal genes on the p arm of chromosome 9 (Wang, 2012). Recently a 5.8 kb deletion was identified in the non-coding space between exons 1 and 2 in the *ABO* gene (Sano, 2015). Yet despite decades of studies of the *ABO* gene, including in large-scale population specific screens allowing for rare variant discovery (i.e., exome and whole genome sequencing), SV has not been previously reported in the coding portion of the *ABO* locus. Here we apply two read-depth (RD) approaches to the analysis of a large exome dataset, and we report the discovery of SV in the *ABO* gene.

4.2 **Methods**

We used *ABO* coding sequence data derived from the NHLBI Exome Sequencing Project (ESP) and the Minority Health *Genomics and Translational Research Bio-Repository Database (MH-GRID)* (NHLBI-ESP, 2016; MH-GRID, 2016). Through the ESP, 15,336 genes were sequenced at high coverage (median depth >100x) in a total of 6,515 unrelated European American (EA, n=4,298) and African American (AA, n=2,217) individuals from 19 different cohorts (Fu, 2013). Through MH-GRID, exome sequences at similar high coverage (median depth >100x) were obtained from a total of 1,313 unrelated AA individuals. The library construction, exome capture, sequencing, mapping, calling, and filtering were carried out as described previously (Johnsen, 2013). Exome sequence data were aligned to NCBI human genome reference GRCh37.

Sequence data were aligned to *ABO* reference sequence obtained from GenBank (NP_065202.2). Average sample read depth for the *ABO* gene was 77x (ranging from 10 to 374x) and includes the entire coding sequence and exon-intron boundaries. Of the 4,298 EA ESP participants, 3,405 had minimum coverage of 50x across the entire targeted *ABO* region. Additionally, 3,027 self-identified African Americans (AA) from both the ESP and MH-GRID were included in our *ABO* analysis bringing our total to 6,432 participants, roughly divided equally between EA (53%) and AA (47%).

For SV discovery, we applied two algorithms optimized for exome datasets. The first, eXome Hidden Markov Model (XHMM) (Fromer, 2012), uses principal component analysis (PCA) and a hidden Markov model (HMM) to detect and genotype structural variants normalized for RD data across samples. Applying this method to the RD obtained with the genome analysis toolkit (GATK), we called small (1–100 kb) SV in our combined set of 6,432 participants (McKenna et al., 2010). We also implemented an orthogonal read-depth based algorithm (CoNIFER) to call SV (Krumm, 2012). CoNIFER uses singular value decomposition (SVD) to detect rare SV and genotype these from exome sequencing data. Both algorithms were run with default parameters. As SV callers are subject to false positives (Teo, 2012), we explored only SV events found in two or more participants and filtered on allele balance (70% cutoff), while allowing for a 50% overlap between unique independent SV (Figure 1A).

Putative novel SVs (identified by both approaches) were further confirmed by TaqMan Copy Number Assays (Dennis, 2012). To accomplish this, a probe specific to *ABO* exon 7 (assay ID Hs01862499_cn, ThermoFisher Scientific, Grand Island, N.Y.) was used to evaluate the copy number of genomic DNA targets using Applied Biosystems real-time system and the Ribonuclease P RNA component H1 (RPPH1) gene as the control assay. To assess copy number, 2.0 μl sample DNA (5ug/mL), 5.0 μl copy number specific TaqMan master mix, 0.2 μl *ABO*, exon 7, assay ID Hs01862499_cn probe, 0.5 μl of copy number specific TaqMan reference assay (RPPH1), and 2 μl PCR quality water

(10.0 μl per sample). African ancestry samples with and without the predicted SV in exon 7 and a HapMap reference sample (NA-12878) without the *ABO* SV were used as controls. Copy number was then calculated using Applied Biosystems Copy caller version 2.0. We then examined nearby duplicated elements annotated in the UCSC hg19 genome browser segmental duplication (SD) and Repeat Masker track's that associated with the theorized exon deletion breakpoints discovered in both our combined exome sequence dataset and a publically available orthogonal NGS dataset called the 1,000 genomes project (1KG). After identifying putative breakpoints using both the repeat masker and SD tracks in the 1KG dataset we then designed long-range PCR primers in the predicted breakpoint region and accurately confirmed the size of our AA specific *ABO* deletion finally calculating it's population frequency.

4.3 **Results**

Using XHMM, 16 putative SVs were predicted in the coding sequence of *AB*O in 6,432 individuals (12,864 chromosomes). Five predicted *ABO* SVs were identified in two or more individuals; four deletions were detected in 32 individuals, and one duplication detected in two individuals. Eleven predicted SVs were found only in a single individual, i.e., singletons (Supplemental Table 1, Figure 1). Using CoNIFER, we identified three predicted *ABO* SVs (a single deletion discovered in five individuals) and two singletons (both deletions including exon 7) (Supplemental Table 1). We focused on SVs found in more than one individual since singletons have a higher rate of false positive detection (Mills, 2011). The five predicted SVs found in 2+ individuals were overlapping and included a putative deletion of *ABO* exons 5-6, a deletion of *ABO* exons 4-7, a deletion of *ABO* exons 5-7 (Figure 1A), and a potential duplication of *ABO* exons 5-7 identified by XHMM. Visual inspection of the RD and allele balance revealed that the putatively identified deletion of *ABO* exons 5-6 also included deletion of exon 7 (Figure 1B) and that the deletions of *ABO* exons 4-7 likely only affected *ABO* exons 5-7. Thus, after visual inspection, all of the putative *ABO* exon deletions detected by XHMM appeared similar

to the *ABO* exon 5-7 deletion identified by CoNIFER. This highlights the importance of manual curation and the benefit of applying more than one read depth based algorithm to robustly predict SV from NGS datasets. It should be emphasized that the RD algorithms applied here should be used as screening tools to detect an SV signal, as many of our initial findings were eliminated during the validation/confirmation process.

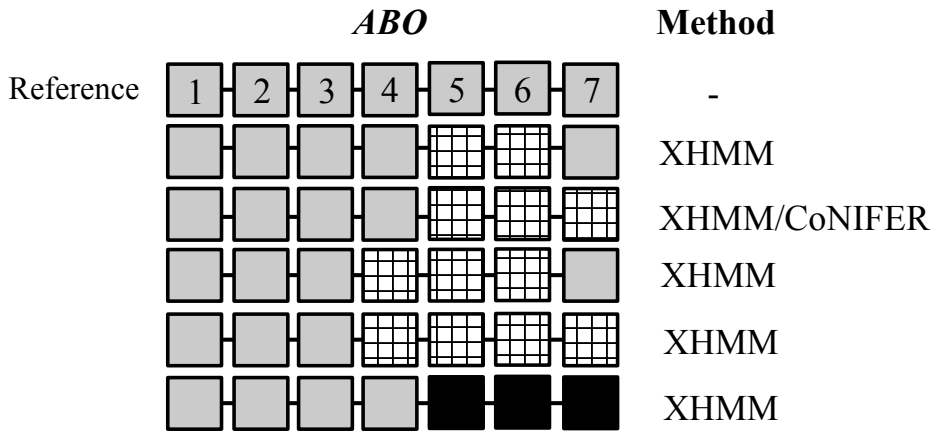

Figure 4.1A: **Putative SV discovered in *ABO* in 6,432 exomes using XHMM and CoNIFER.** Putative SV discovered in *ABO* in 2+ individuals using two read depth methods are shown; each gray box represents intact (2 copies of) exons in the *ABO* gene. Although *ABO* is a negative strand gene it is shown with left to right orientation. Gridded boxes represent exons predicted to be deleted using both XHMM and CoNIFER. Black boxes represent *ABO* duplications identified using XHMM. Both read depth algorithms putatively predicted deletion of *ABO* [E5-7].

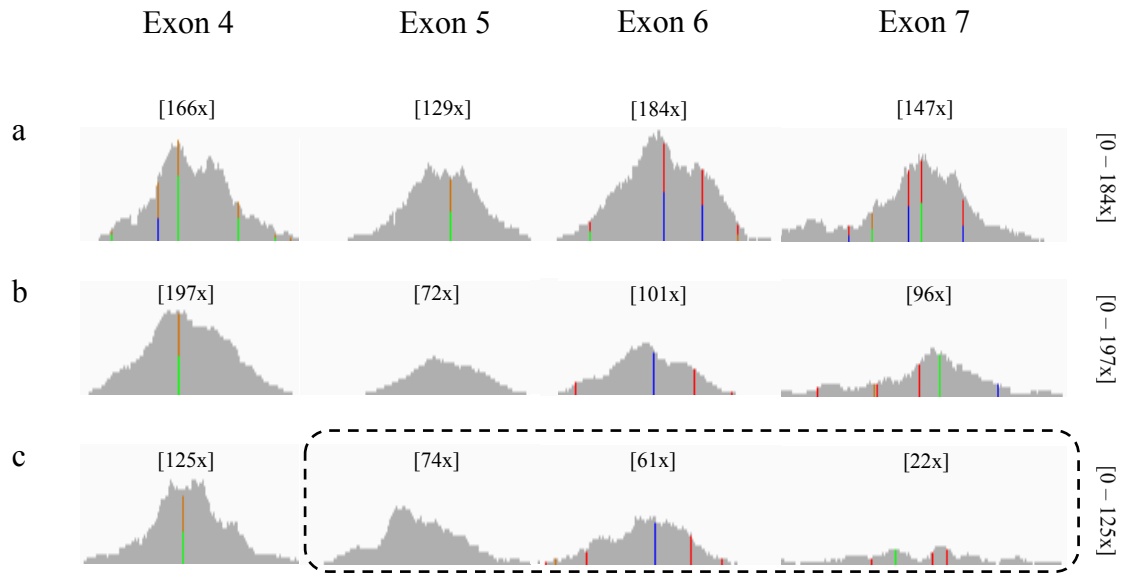

Figure 4.1B: **Examples of read-depth data in support of predicted SV in *ABO*.** Integrative genomics viewer (IGV) read depth comparisons are shown for *ABO* exons 4-7. *ABO* is a negative strand gene however each exon is displayed from left to right; Colored vertical bars in IGV indicate the number of reads supporting each allele for single nucleotide variants in each exon (i.e., allele balance.) Loss of heterozygosity and lower read depth (i.e., mono-colored vertical line or solid grey vertical line) is consistent with a deletion in a given exon, versus heterozygosity and higher read depth (i.e., a multi-colored vertical line) of a SNV is consistent with two copies of an exon (e.g., exon 4). (a) An individual with average read depth at *ABO* (exons 4, 5, 6, and 7). (b) An individual predicted to carry a deletion of E5-7 by both XHMM and CoNIFER. (c) An individual predicted to harbor a putative deletion of E5-6 and not exon 7 by XHMM. However, visual inspection of read depth and allele balance clearly show low coverage consistent with a deletion of exon 7 as well. Dashed region represents minimal size of the putative deletion.

Using a TaqMan probe in *ABO* exon 7, we tested 16 individuals predicted to carry one copy (deletion), two individuals predicted to have three copies (duplication), and eight individuals predicted to have two copies of exon 7 (control) in the *ABO* gene. While a deletion affecting exon 7 was confirmed in all individuals (n=16), the *ABO* duplication in exon 7 (E5-7) could not be confirmed and, therefore, represents a false positive (Supplemental Figure 1). Because whole exome sequencing only targets exonic sequence (exclu-

ding intronic regions), our validation experiments indicated that one of the events identified in 2+ individuals was larger than the length predicted from our exome-based XHMM analysis (i.e., ([E5-6 deletion]). By performing a combined analysis (integrated SV call-set) with visual inspection, we were able to refine our call boundaries and robustly predict and validate SV in the *ABO* gene. In addition to the recurrent deletions, we observed rare lower confidence overlapping singleton deletions encompassing exons 5 to 7 with variable proximal and distal breakpoints. As DNA was not available for validation, we have excluded these events from our allele frequency estimate for the AA *ABO* deletions (Supplemental Figure 2).

4.4 Discussion

To our knowledge, SV in the coding portion of the *ABO* gene has not been previously suspected or reported. Interestingly, the partial *ABO* gene deletion we identify has a greater than 50% reciprocal overlap with an event identified in the third and final stage of the 1,000 Genomes Project (1KG) reference panel (Sudmant, 2015). In further support of our discovery of SV in *ABO*, our 3-exon deletion ([E5-7 deletion]) including the distal exon 5 breakpoint (Figure 1B) was identified in both our cross-sectional analysis of a combined exome dataset and the phase 3 1KG (Sudmant, 2015). Finally in both the 1KG and our combined exome dataset, the 3-exon *ABO* deletion was exclusively found in individuals of both African and African admixed ancestry with a similar allele frequency; our combined exome dataset (AA 0.00469) and 1KG phase 3 (African (0.0129), admixed American (0.004)). The slight differences detected in SV allele frequency likely reflect the differences in historical admixture in African, African American, and admixed American populations (Parra, 1998).

Further examination of the recurrent deletions identified a segmental duplication (SD) associated with the distal breakpoint region. Previous studies have suggested that SDs are capable of driving increased CNV mutation rates, not only, through flanking

direct orientation repeats (non-allelic homologous recombination**),** but also through a replication/repair based mechanism in regions adjacent to duplications (Itsara, 2010). To this end we examined all nearby duplicated elements annotated in the UCSC hg19 genome browser SD and repeat masker track that associated with the theorized exon 5-7 deletion breakpoints discovered in the whole genome 1KG phase 3 data. Both breakpoints reported in the phase 3 1KG data were mapped to intronic regions (proximal break point between exons 4-5 and distal break point between the end of exon 7 and *ABO*'s neighboring gene, *OBP2B*). Through our repeat masker analysis we identified two flanking Alu retrotransposons within the same sub-family (AluSx1) with direct orientation on either side of the three exon deletion suggesting that these AluSx1 mobile elements may be the mechanism responsible for these deletion alleles. We designed long-range PCR primers with ~200 bp flanking buffer from our theorized breakpoints in an effort to confirm the presence of the three exon deletion in AA samples identified via our combined RD analysis. Our results confirm a deletion impacting exons 5, 6, and 7 of *ABO* (Supplemental Figure 3) and could be consistent with a deletion mechanism likely involving the breakpoint flanking AluSx1 elements.

By using sequence data from thousands of AA exomes, we were able to characterize, validate, and provide population frequency estimates for an SV, which results in a partial *ABO* gene deletion. This finding strongly supports the utility of exploring all types of genetic variation (e.g., SV) in large emerging datasets. Moreover our study shows it is possible to identify previously unrecognized sources of variation, in this case a SV in the *ABO* gene leading to a novel predicted LOF O allele. This sets precedence for other yet unidentified *ABO* alleles that could include SV (i.e., unidentified *ABO* deletions and duplications) in diverse populations. A more comprehensive understanding of variation in blood group genes in minority populations has the potential to refine our knowledge of ABO as a major genetic modifier of diseases including heart attack, stroke, and thrombosis. The approaches presented here can be readily applied to many other genes that direct expression of blood cell antigen systems, including those with known SV (e.g.,

Rhesus blood groups (RH), MNS (glycolphorins), major histocompatibility complex (MHC) antigens, minor histocompatibility antigens (mHAs), killer cell immunoglobulin receptor genes (KIRs)) and those not known to harbor SV, such as was previously thought for *ABO* (Storry, 2009; Mullalay, 2007).

(PAGE INTENTIONALLY LEFT BLANK)

Chapter 5

CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter, I present future ideas for ABO blood group genomics, including increasing the accuracy and the potential for the adoption of sequence-based blood typing methods in translational research and clinical settings. I describe how the combination of serological and NGS-based systems (e.g., ABO-Seq) might allow clinicians to better identify atypical ABO types to enhance transfusion and organ transplantation practice, and how refined ABO typing could be applied in association studies for common complex disorders. I discuss the importance of exploring the function of *ABO* variation of unknown significance, particularly in forthcoming, large-scale population-based analysis. Lastly, I discuss the need for characterizing the complete spectrum of human genetic variation, especially for underrepresented populations and its importance to the future of genomic-based medicine.

5.1 Summary

Although *ABO* is one of the most well-studied loci in the human genome, in this dissertation I have shown that opportunities for novel discovery remain. I believe ABO and blood-group systems, in general, present unique opportunities to study clinically relevant human traits at both the genotypic and phenotypic levels. Common antigens in blood-group systems are routinely typed in the clinical settings through serology-based methods. While serology is the gold standard for determining blood type, in this dissertation I have shown that sequencing platforms are capable of determining ABO blood type and can provide high resolution characterization of genetic variation at the *ABO* locus (Chapter 3). For this reason, I believe these tools should be useful to complement serological methods. However, the broad use of sequence-based determination of ABO blood type currently has several limitations. There are significant limitations when translating described rare variation with corresponding serology (BGMUT alleles)

to genomic coordinates and *vice versa* (relating sequence-based variation to BGMUT alleles). I discuss this in greater detail in section 5.2 of this chapter.

At present, sequence-based-blood typing is reliant on curated DNA variant data with corresponding serologic data to predict variant function, although this implies that sequence-based typing is not appropriate as a stand-alone method in the context variant functionalization or in clinical settings. The pairing of NGS methods and serology makes *ABO* a good model for understanding loss of function variation in the human genome. For example, the majority of loss of function in the *ABO* gene is the result of a single deletion in exon 6 occurring on the A blood type haplotype background, but it is not the exclusive cause of loss of function in *ABO*. The application of NGS and serology is a powerful way to discover novel ways that can cause a loss of function in *ABO* (see section 5.3).

From an evolutionary perspective, the genetic heterogeneity in *ABO* is interesting because it likely represents our evolutionary history. The heterogeneity in the *ABO* locus may reflect the selective pressure of various pathogens on our ancestor's genomes. This genetic variation is manifested at the phenotypic level in the form of blood group antigen diversity. Previous studies measuring ABO antigen diversity via serology have detected variation across all human populations (Cavalli-Sforza, Barrai, and Edwards, 1964). In this dissertation, I have re-characterized the distribution of genetic variation in *ABO* based on NGS-derived genotypes (Chapter 2). I discuss my findings in section 5.7 below.

Finally, throughout my dissertation, I have made mining human genetic variation in underrepresented minority populations a priority. The genomics community has largely generated sequence data from individuals of European ancestry. Consequently I was able to identify previously unsuspected structural variation in one of the most well-studied genes in the human genome (Chapter 4). A comprehensive understanding of variation in blood-group genes in minority populations has the potential to (1) refine our knowledge

of genotype-phenotype correlation in intermediary biomarkers associated with cardio-vascular disease and (2) create a more comprehensive understanding of health disparity in the U.S. I discuss this topic in detail in sections 5.8 and 5.9.

5.2 Translational challenges

A major challenge that will hinder the application of sequence-based ABO typing is the translation of allelic variation cataloged in the BGMUT database into genome-based coordinates. Many *ABO* variants lack the standard annotation formats (e.g., HGVS, CDS, ISBT nomenclature, and Sanger-sequencing entries). In my attempts to convert BGMUT *ABO* alleles to genomic coordinates, many of the BGMUT entries did not correspond to the allele that was entered in the dbSNP database. Moreover, for some alleles the specified variant did not correspond to the predicted amino acid change specified in the entry and were also unavailable in dbSNP. As a result, only BGMUT entries with accompanying sequence data could be reliably translated to genomic coordinates in an automated fashion. This reduced the number of alleles in the BGMUT database useful in this analysis by half (151/377) and excluded cataloged variation with valuable, corresponding serological data. Manual curation of the remaining alleles could define additional *ABO* genetic variation useful in predicting ABO phenotypes — this is important in bringing the blood-group typing into the genomic era.

According to a recent analysis of variants included in clinical variation databases, there is a substantial discordance between both annotation tools (SnpEff, Variant effect, predictor, and variation reporter) that generate transcript and protein-based variant nomenclature from genomic coordinates according to guidelines by the Human Genome Variation Society (HGVS). In fact, Church et al. report discordance in the range of 50-70% for protein variants, numbers that are not adequate for clinical typing. These results reveal significant inconsistencies between variant representation across tools and databases. Additionally, these results highlight the urgent need for adoption and

adherence to uniform standards in variant annotation, with consistent reporting on the genomic reference, to enable accurate and efficient data-driven clinical care (Church et al., 2016).

5.3 Variant prediction challenges

The importance of genotype-phenotype information is also highlighted by the limitations of predicting variant function *in silico*. Sequence-based computational tools are capable of estimating the potential effect of a rare variant (i.e., singleton or VUS) on structure or function, such as GREP, POLY-PHEN2, and on CADD scores; however, although most common group O haplotypes share a common, crippling polymorphism, a growing number of described alleles feature other polymorphisms that render their protein nonfunctional (Hosseini-Maaf et al., 2005). These alleles are similar enough to the consensus A haplotype; if a sequence-based algorithm was not programed to detect all possible non-deletional O haplotypes, an errant phenotype could be predicted from the genotype (Yazer, Hosseini-Maaf, and Olsson, 2008a). Some of these non-deletional O alleles might actually encode a protein with weak and variable A-antigen ability (Yazer et al., 2008b). As a result, accurate prediction of LOF in *ABO* remains a serious issue and thus transitioning to ABO blood type based exclusively on sequence is highly unlikely. This difficulty highlights the importance of current efforts (Bloodwork's Northwest Blood-seq project) to catalog variation in ABO by NGS sequencing and the importance of providing corresponding serological phenotype data. The future of blood-group genomics will involve complementing blood-group serology with genome screening tools to understand both typical and atypical serology.

5.4 Functionalizing rare variation in *ABO*

Through the analysis of large-scale NGS datasets, I have identified variation present in the coding portion of the *ABO* gene. The majority of variants are rare and are the

result of SNVs, indels, and SVs that I predict affect the structure or function of the *ABO* glycosyltransferase. However, many of the variants identified in these large-scale screens of human populations are variants with unknown significance (VUS). Heuristics such as CADD, GERP, and others attempt to predict variant function but they do not truly determine the effect VUS will have on structure or function. Recent advances in genome-editing technology (e.g., CRISPR-Cas9) will permit functional validation of VUSs in human-specific cell lines (Starita et al., 2015). With genome editing (Doudna and Charpentier, 2014), one could perform targeted mutagenesis of (1) predicted LOF/VUS and (2) missense variation in the coding portion of *ABO* to functionalize this variation in human cell lines. To produce accurate LOF frameshifts, one could employ a lentiviral-based CRISPR/Cas9 homology-directed repair approach to edit the *ABO* gene *in situ* in cell lines. An antibiotic-resistance marker incorporated in the viral vector might allow selection for transduced cells, which are more likely to have the desired LOF edits. Selected single cells could also be sorted into a 96-well plate using flow cytometry; these cells grow into clonal populations, and screened to identify successfully edited cell lines, could assess function. In the future I imagine simultaneously assessing variation across multiple-phased loci, i.e., saturation mutagenesis or massively parallel *ABO* haplotype editing (Findlay, Boyle, Hause, Klein, and Shendure, 2014). Ultimately, selecting a cell model that is capable of displaying A, B, and H antigen structures is necessary for accurate functional validation of ABO haplotypes.

As human RBCs do not contain DNA, one could select an early-stage RBC such as a proerythroblast cell line that contains DNA for editing purposes to use homology-directed repair to edit in *ABO* variant libraries (i.e., *ABO* haplotypes). I would likely start with a handful of known haplotypes and then expand library construction to combinations and permutations of *ABO* haplotype variation including rare variation. Next I would attach fluorescently labeled Anti-A and Anti-B antibodies to our newly edited RBCs and use fluorescence cell analysis to separate them into A, B, AB, and O categories (Noderer

et al., 2014); I would then perform targeted NGS sequencing of the *ABO* locus (O'Roak et al., 2012) to determine which ABO haplotypes correspond to the antibody selection.

I would also attempt to use a phage display or yeast cell to construct a glycan display model. This type of deep mutational scanning system would help determine which rare or novel haplotypes lead to LOF, weak alleles, or altered donor glycan specificity (such as Cis-AB). Moreover, these systems will aid in validating predictive heuristic algorithms such as CADD and GERP (Kircher et al., 2014; Cooper et al., 2005). By functionalizing haplotype variation in the *ABO* gene using massively parallel genome editing and FACS to sort those edited cells, the broader field will then have a system to functionalize rare *ABO* variants *in vitro*.

5.5 Known associations, outcomes, and intermediate phenotypes: directionality

ABO has been associated with various disease phenotypes, including thrombosis and vascular disease (Wiggins et al., 2009). However, many of the initial studies have used noncoding SNVs present on genome-wide arrays or imputed from HapMap as tags/ proxies for functional *ABO* SNVs (A1, A2, B) or the O c.261delG (Campos et al., 2011). These proxies are not in perfect linkage disequilibrium (LD), even in individuals of European ancestry; for individuals of African ancestry, LD has yet to be comprehensively described (Johnson et al., 2008). Therefore, it is possible that ABO haplotypes derived from NGS data could provide new insights into previous associations with *ABO*.

In addition to previous associations with CVD, *ABO* genotype is also the major genetic modifier of both von Willebrand factor (VWF) and factor VIII (FVIII) plasma concentrations (Zabaneh et al., 2011). Increased plasma concentration levels of VWF and FVIII are associated with a risk of myocardial infarction and cardiovascular disease (CVD) (Smith et al., 2010). Normal adult subjects with type O blood have approximately 30% lower levels of VWF and FVIII than those with non-O blood types (Trégouët et al.,

2009) and are more likely to have the bleeding disorder von Willebrand Disease (Johnsen and Ginsburg, 2015). FVIII levels are largely dependent on VWF, as VWF binds and protects FVIII from proteolysis in circulation, and the correlation between ABO and FVIII is most likely mediated by the influence of ABO on VWF (Smith et al., 2010). For circulating levels of intracellular adhesion molecule 1 (ICAM-1) associated with myocardial infarction (MI), as well as CVD risk, the A1 subtype has been associated with decreased levels of ICAM-1 (Paré et al., 2008). By exploring the relationships between common and rarer haplotypes in *ABO*, future research efforts could provide some insights into associations between quantitative traits (VWF, FVIII and ICAM-1) associated with *ABO* and hemostasis, thrombosis, and vascular disease susceptibility.

5.6 ABO subtypes and cardiovascular disease

The large-scale analysis of *ABO* haplotypes also creates a path forward for future efforts to explore the genotype-phenotype relationship between individual *ABO* sub-type haplotypes such as A2 versus A1, intermediary biomarkers, and early-onset vascular events. This work could also potentially expand analyses to other haplotypes/subtypes found in the *ABO* region. I hypothesize that individuals with *ABO* subtype/haplotype A2 will have a decreased association for MI and stroke when compared to non-A2 subtype individuals, with the exception of O blood type individuals who will have the lowest association with both MI and stroke. Moreover, because intermediate biomarkers have also been associated with CVD, I hypothesize that, in addition to lower VWF levels, A2 subtype individuals will have lower FVIII and VWF levels when compared to non-O and non-A2 individuals (Sousa, Anicchino-Bizzacchi, Locatelli, Castro, and Barjas-Castro, 2007). Others have demonstrated that the O blood type has a 67% lower risk of venous thrombosis (VTE) when compared to non-O blood groups (Trégouët et al., 2009). Additionally, the A2 blood group has a 47% lower risk of VTE when compared to other non-O blood group phenotypes (Trégouët et al., 2009). Finally, in a recently published

GWAS involving 1,503 VTE patients, blood type A2 was shown to be associated with lower VTE risk than other non-O blood groups (Trégouët et al., 2009). The application of sequence-based ABO subtyping technology can be applied to refine the CVD disease associations and explore these associations even further (Chen, Yang, Xu, and Li, 2016).

Furthermore, very little is known about rare variation in *ABO* and its association with disease in African Americans who are disproportionately affected by cardiovascular disease when compared to individuals of European ancestry in the U.S. (Feinstein et al., 2012). Specifically, identifying population-specific haplotypes and predicting novel weak serological phenotypes (such as haplotypes discovered exclusively in minority populations) are promising candidates for molecular/functional validation that might have biological significance.

5.7 Clues from our past: ABO blood type and infectious disease susceptibility

Disease association with ABO is not limited to CVD. For example, individuals with the A blood type are at a higher risk of developing several types of cancer, such as some forms of pancreatic cancer and leukemia (Wolpin et al., 2010). Non-O blood-type individuals are also more prone to malaria (Fry et al., 2008). Experimental studies manipulating ABO glycans on erythrocytes indicates that macrophages have enhanced uptake of infected red blood cells if those blood cells are type O, providing a potential mechanism by which blood group O is protective in malaria (Wolofsky et al., 2012).

Additionally, links have been made between ABO blood type and norovirus infection. Norovirus is a highly contagious gastrointestinal pathogen often responsible for vomiting and diarrhea on cruise ship vacations, in institutional settings such as nursing homes, and in community gastroenteritis outbreaks. It is well known that A, B, and H antigens are expressed on endothelial cell surfaces including the intestinal lining of the gut (Yamamoto, Cid, Yamamoto, and Blancher, 2012). Noroviruses attachment appears to

be specific to the presence of ABO and/or H antigens expressed on cell surfaces in the gut. Interestingly, specific strains of norovirus bind to various antigens of the ABO, H, and Lewis groups (a related carbohydrate group) differently depending on the strain of norovirus. It is postulated that each strain of norovirus has proteins that are adapted to attach tightly to specific A, B, or H antigens, but not others (Huang et al., 2003). This might explain why an individual's ABO blood type can influence norovirus-infection susceptibility (Huang et al., 2003).

While my analysis of *ABO* haplotype diversity did not focus on modes of selection shaping diversity in the *ABO* locus or demographic migratory history, I believe that detailed analysis of *ABO* haplotypes among diverse populations promises to create a more refined version of the migratory history of both contemporary and ancient hominids *via* the analysis of rare variation in the *ABO* gene, especially in indigenous populations. Indigenous populations offer new keys into understanding our history of co-evolution with pathogens (Tishkoff, 2015). Refining our understanding of *ABO* haplotype variation in contemporary human populations, ancient hominids, and other known primate species offers keys to unlocking the evolutionary history of trans-species polymorphisms (Ségurel et al., 2012), a more comprehensive understanding of balancing selection (Charlesworth, 2006), and potentially illuminates our history of interaction and co-evolution with pathogens.

I anticipate an explosion of research connecting blood-group gene diversity and gut-microbiome diversity that could lead to a refined version of the history of *ABO* and microbiome diversity and interaction. The future of *ABO* population genomics might include enhancing the resolution of analyses focusing on the interaction and co-evolution of blood-group antigen structures and human gut micro-biome diversity. Another direction blood-group genomics might take in the future is the analysis of complex variation in multiple blood groups (e.g., *ABO*, *RHD*, *RHCE*, *GYPA*, etc.) in diverse populations in an attempt to evaluate for evidence of selective evolutionary

pressures in both common complex disease and infectious-disease susceptibility. These aggregate multi-blood-group gene-family diversity analyses might offer additional strategies for researchers to trace primate-pathogen interaction history and to assess directional modes of natural selection as forces that shape diversity in the human genome, e.g., positive and balancing selection. As these perspectives are based on ABO and balancing selection, they could be extended to HLA analysis as well, for example, the effects of smallpox exposure *via* European contact on HLA genotype diversity among both contemporary and ancient indigenous populations in the Americas (Lindo et al., 2016).

5.8 Why genetic research must be more diverse

One theme that is present in each chapter of this paper is the importance of exploring diverse populations for human genetic variation. In Chapter 2 I focused on resolving *ABO* haplotype diversity in ~30 distinct populations and two ancient hominids. Chapter 3 focused on the development of a new method to subtype blood in three distinct populations, highlighting the importance of rare population-specific variation in the coding portion of *ABO* as a strategy to improve our ABO-seq algorithm. In Chapter 4, by focusing our efforts on minority populations (i.e., individuals of African ancestry), I was able to identify novel LOF SV in one of the most well-studied genes in the human genome. Moreover, a comprehensive understanding of variation in blood-group genes in minority populations has the potential to (1) re-define the information used to optimally match donors and recipients in transfusion and transplant therapies, (2) refine our knowledge of genotype-phenotype correlation in intermediary biomarkers associated with vascular disease, and (3) create a more comprehensive understanding of health disparity in the U.S. Throughout this thesis I have focused on the identification of both common and rare variation in the *ABO* gene; but the approaches presented here can be readily applied to many other genes that direct expression of blood-cell antigen systems,

including those with known SNVs, indels, and SVs (e.g., Rhesus blood groups [RH], MNS [glycophorins], major histocompatibility complex [MHC] antigens, minor histocompatibility antigens [mHAs], and killer-cell immunoglobulin receptor genes [KIRs]), as well as those loci not known previously to harbor SV, such as we reported in *ABO*.

Over the past decade, researchers have dramatically improved our understanding of the genetic basis of complex chronic diseases — for example, cancer, cardiovascular disease, Alzheimer's disease, and type-2 diabetes — through more than 1,000 genome-wide association studies (GWAS). Yet common variation identified with specific diseases generated through these GWAS studies have had far less impact on the world's population as a whole since 96% of the subjects included in GWAS as of 2011 have focused on individuals of European descent (Bustamante, Burchard, and la Vega, 2011). Moreover, it is well known that humans metabolize drugs differently based on both common and rare variation in many drug-metabolizing genes (e.g., VKORC1 [Rieder et al., 2005; Nelson et al., 2012]). Yet as of 2015, 95% of subjects included in clinical trials, to test new drugs have focused on individuals of European descent (Oh et al., 2015) despite minority populations making up an estimated 45% of the U.S. population (Reardon et al., 2015). While clinical trials fail to include these underrepresented populations, minorities are affected in disproportionately greater number by every metric used to evaluate health in the U.S., including higher infant mortality rates, shorter average life spans, and higher rates of the first, second, and third causes of death in the U.S.: cardiovascular disease, cancer, and type-2 diabetes (Oh et al., 2015).

Much of the health disparity observed in the U.S. is intertwined by a relationship between socioeconomic status (i.e., environment) and genetic predilection to common complex disease (Florez et al., 2009; Reardon, 2015). However, there is hope as NGS efforts such as the NHLBI-ESP, MHGRID, the 1,000 Genomes Project, ExAC, and most recently TOPMed, have prioritized the inclusion of diverse populations in large-scale screens in an attempt to create a comprehensive understanding of rare variation in the

human genome. As of 2015, the Precision Medicine Initiative (PMI) cohort organization program has pledged to overrepresent minority populations in the study relative to the share of minorities in the U.S. (Reardon, 2015). By overrepresenting the number of minorities included in the PMI researchers may be able to draw statistically significant conclusions about small groups of underrepresented minorities and the genetic mechanisms that contribute to the disproportionate trends we observe in disease (Burchard, Oh, Foreman, and Celedón, 2015). Finally, one key contribution to understanding the full extent of diversity in large-scale human genome-sequence studies in the future will be the development and assembly of population-specific genome references (Evan Eichler, personal communication; Fakhro et al., 2016).

5.9 Aiming for ethnic balance in large-scale genomic studies in the future

Alcoholism is s health issue in Native American communities, for example; a study such as the PMI could help reveal genetic and environmental factors that might underlie this susceptibility (Reardon, 2015). However, Native Americans make up just 1.6% of the total U.S. population and if represented proportionally in the PMI cohort program of one million individuals it would require the recruitment of 16,000 Native Americans. Additionally, focusing on sub-groups (e.g., Navajo, Lakota, Tulalip, etc.), socio-economic status, age, etc., would further reduce this sample's size (Reardon, 2015).

Recruiting and retaining members of underrepresented communities represents a huge challenge moving forward with large-scale population-based screens. There may not be the resources to find recruitment information online; in addition, these populations may lack access to mobile or smart phone apps (i.e., mobile health interventions) and tend to have a general distrust of the U.S. government as a result of the historical exploitation of their communities (Yong et al., 2016). To quote Esteban Burchard at the University of California, San Francisco Medical School, "Just because you can study patients at an ivory-tower academic institution doesn't mean you can do it in rural

Appalachia." I welcome these challenges and I hope to engage some of these groups in the next phase of my scientific career. I believe that the future of underrepresented minority recruitment and engagement in large-scale genome-sequencing studies will involve significant advances in community-based participatory-research methods (Morales et al., 2016; Woodahl et al., 2014). Moreover, in order for the democratization of genome-sequencing studies to occur we will need significant technological advances that include (1) access to the internet in remote communities, (2) remote access to passively parallel computation, (3) point-of-care or mobile genome sequencing, and (4) community-based education tools that will allow for the evolution of informed consent (McCaughey et al., 2016). Once these tools are in place the genomics community can begin to truly reimagine their relationship with remote/rural communities and begin to include them in the future of predictive and preventative medicine.

5.10 Closing thoughts

ABO is one of the most well-studied genes in the human genome (F. Yamamoto et al., 2012). The efforts highlighted in this thesis are an attempt to translate greater than a century's worth of research on the *ABO* gene from a genetic perspective to a genomic perspective. The contributions and discoveries highlighted are a modest addition to the existing body of work on genetic variation in *ABO*. I have been both humbled and inspired to learn about the spectrum of human genetic variation that exists in our genomes though analysis of a single loci, *ABO*. *ABO* is an elegant example of a loci with extensive heterogeneity that has both direct clinical impact (i.e., blood transfusion therapy) and indirect clinical impact that has just begun to reveal itself to the biomedical research community (i.e., CVD and cancer susceptibility). The extensive heterogeneity we observe in *ABO* today is a signature of natural selection, a powerful reminder of the diversity that exists in the human genome in global populations as a result of local adap-

tion to pathogens found in environments where our primate ancestors have subsisted for millennia.

While many other locally adaptive traits exist in the human population (*MCM6* and lactose tolerance, *DARC* and malaria/sickle cell anemia, etc.), none has been as well known to the non-technical public as the *ABO* gene. Because of ABO's popularity as a trait with measurable diversity among modern human populations, popular science books have been published and have even become national best sellers with little scientific rigor to support them. For example, *Eat Right for Your Blood Type* claims that, "Your ABO blood type might play a role in personalizing your diet" (D'Adamo et al., 1997). While this assertion has no scientific backing it highlights the general public's interest in what makes us different as human beings, our genomes. However, while it is often difficult to translate the value of genomics and genome sequencing to the non-technical public, the *ABO* blood groups remain a wonderful example to use when explaining how diversity in the human genome manifests in measureable human characteristics that we can observe.

While predicting actual ABO blood type from sequence alone may not be adopted into routine clinical practice in the near future, resolving *ABO* haplotypes from next-generation sequence data represents a real success story for the transfusion medicine community in the precision medicine era (Johnsen, 2015). By complementing serologically determined ABO blood type we can solve challenging clinical cases, and ABO-incompatible organ-transplant successes and failures demonstrate the need for higher resolution studies of ABO and its subtypes. While the impact of sequence-based ABO blood-group typing is still in its preliminary stages, there is no doubt that at least some patients will be positively affected by refining blood group typing to decipher ambiguous atypical serological samples. Similar promise exists for screening and testing for a variety of other blood group genes, including the Rhesus blood groups (RH), MNS (glycolphorins), major histocompatibility complex (MHC) antigens, minor histocompatibility

antigens (mHAs), and killer-cell immunoglobulin receptor genes (KIRs). There is, however, much work yet to be done.

Finally, while the PMI might offer predictive and preventative medicine to those who choose to participate. I am reminded of a question I received from an elderly, Native American man at the National Museum for the American Indian a few years ago while giving a talk. He asked me, "Why should I care about genome sequencing?" He added, "Genome sequencing? We haven't had clean water on our land since 1912!" At the time I was stumped. On any given day, half of the world's hospital beds are occupied by patients suffering from illnesses associated with lack of access to safe water and lack of sanitation (Brown, Neves-Silva, and Heller, 2016).

The truth is, there is no single or simple solution to the issues he raises. However, I take his comment as a reminder of the inequalities that exist in the way health impacts quality of life, and in healthcare delivery around the globe. The unfortunate reality is that all too often, underrepresented minority communities are excluded from research that seeks to address the severe health issues they experience at disproportionate rates. In order to improve health outcomes, we must include these underrepresented communities in the future of precision medicine. I believe that genome sequencing is an important part of that process. It is up to the next generation of genome scientists to find the right approach to reducing health disparities to ensure that the benefits of predictive and preventative medicine are shared by all.

When you see a giant wave moving in your direction don't brace for impact; dive into the wave head first with all of your strength and come out the other side unscathed.

Ancient Hawaiian proverb

(THIS PAGE LEFT INTENTIONALLY BLANK)

REFERENCES

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [Et Al.]*, *Chapter 7*, Unit 7.20–7.20.41. http://doi.org/10.1002/0471142905.hg0720s76

Amirzadegan, A., Salarifar, M., Sadeghian, S., Davoodi, G., Darabian, C., and Goodarzynejad, H. (2006) Correlation between ABO blood groups, major risk factors, and coronary artery disease. *International Journal of Cardiology*, *110*(2), 256–258. http://doi.org/10.1016/j.ijcard.2005.06.058

Amos, W. (2016) The quantity of Neanderthal DNA in modern humans: a reanalysis relaxing the assumption of constant mutation rate. Biorxiv. doi: http://dx.doi.org/10.1101/065359

Auer, P. L., Johnsen, J. M., Johnson, A. D., Logsdon, B. A., Lange, L. A., Nalls, M. A., et al. (2012) Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *American Journal of Human Genetics*, *91*(5), 794–808. http://doi.org/10.1016/j.ajhg.2012.08.031

Bernstein, F. (1924) Ergebnisse einer biostatischen zusammenfassenden Betrachtung ûber die erblichen Blutstrukturen des Menschen. *Klin Wschr*, 3, 1495–1497.

Björk, S., Breimer, M. E., Hansson, G. C., Karlsson, K. A., and Leffler, H. (1987) Structures of blood group glycosphingolipids of human small intestine. A relation between the expression of fucolipids of epithelial cells and the ABO, Le and Se phenotype of the donor. *The Journal of Biological Chemistry*, *262*(14), 6758–6765.

Bourke, G. J., Clarke, N., And Thornton, E. H. (1965) Smallpox Vaccination: Abo and Rhesus Blood Groups. *Journal of Medical Genetics*, *2*(2), 122–125.

Brown, C., Neves-Silva, P., and Heller, L. (2016) The human right to water and sanitation: a new perspective for public policies. *Ciência and Saúde Coletiva*, *21*(3), 661–670. http://doi.org/10.1590/1413-81232015213.20142015

Browning, S. R., and Browning, B. L. (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews. Genetics*, *12*(10), 703–714. http://doi.org/10.1038/nrg3054

Burchard, E. G., Oh, S. S., Foreman, M. G., and Celedón, J. C. (2015) Moving toward true inclusion of racial/ethnic minorities in federally funded studies. A key step for achieving respiratory health equality in the United States. *American Journal of Respiratory and Critical Care Medicine*, *191*(5), 514–521. http://doi.org/10.1164/rccm.201410-1944PP

Bustamante, C. D., Burchard, E. G., and la Vega, De, F. M. (2011) Genomics for the world. *Nature*, *475*(7355), 163–165. http://doi.org/10.1038/475163a

Cai, X., Jin, S., Liu, X., Fan, L., Lu, Q., Wang, J., et al. (2013) Molecular genetic analysis of ABO blood group variations reveals 29 novel ABO subgroup alleles. *Transfusion*, *53*(11 Suppl 2), 2910–2916. http://doi.org/10.1111/trf.12168

Campos, M., Sun, W., Yu, F., Barbalic, M., Tang, W., Chambless, L. E., et al. (2011) Genetic determinants of plasma von Willebrand factor antigen levels: a target gene SNP and haplotype analysis of ARIC cohort. *Blood*, *117*(19), 5224–5230. http://doi.org/10.1182/blood-2010-08-300152

Carlsten, C., Halperin, A., Crouch, J., and Burke, W. (2011) Personalized medicine and tobacco-related health disparities: is there a role for genetics? *Annals of Family Medicine*, *9*(4), 366–371. http://doi.org/10.1370/afm.1244

Cavalli-Sforza, L. L., Barrai, I., and Edwards, A. W. (1964) Analysis of human evolution under random genetic drift. *Cold Spring Harbor Symposia on Quantitative Biology*, *29*, 9–20.

Charlesworth, D. (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, *2*(4), e64. http://doi.org/10.1371/journal.pgen.0020064

Chen, Z., Yang, S.-H., Xu, H., and Li, J.-J. (2016) ABO blood group system and the coronary artery disease: an updated systematic review and meta-analysis. *Scientific Reports*, *6*, 23250. http://doi.org/10.1038/srep23250

Chou, S. T., Jackson, T., Vege, S., Smith-Whitley, K., Friedman, D. F., and Westhoff, C. M. (2013) High prevalence of red blood cell alloimmunization in sickle cell disease despite transfusion from Rh-matched minority donors. *Blood*, *122*(6), 1062–1071. http://doi.org/10.1182/blood-2013-03-490623

Clausen, H. and Hakomori, S. (1989) ABH and related histo-blood group antigens: Immunochemical differences in carrier isotypes and their distribution. *Vox Sanguinis*, 56, 1–20.

Clausen, H., White, T., Takio, K., Titani, K., Stroud, M., Holmes, E., Karkov, J., Thim, L., Hakomori, S. (1990) Isolation to homogeneity and partial characterisation of histo-blood group A defined Fucα1–2Galα1–3-*N*-acetylgalactosaminyltransferase from human lung tissue. *Journal of Biological Chemistry*, 265, 1139–1145.

Cooling, L. L. W., Kelly, K., Barton, J., Hwang, D., Koerner, T. A. W., and Olson, J. D. (2005) Determinants of ABH expression on human blood platelets. *Blood*, *105*(8), 3356–3364. http://doi.org/10.1182/blood-2004-08-3080

Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7), 901–913. http://doi.org/10.1101/gr.3577405

Crow, J. F. (1993) Felix Bernstein and the first human marker locus. *Genetics* (Vol. 133, pp. 4–7). Genetics Society of America.

Dennis, M. Y., Nuttle, X., Sudmant, P. H., Antonacci, F., Graves, T. A., Nefedov, M., et al. (2012) Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell*, *149*(4), 912–922. http://doi.org/10.1016/j.cell.2012.03.033

Doudna, J. A., and Charpentier, E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096–1258096. http://doi.org/10.1126/science.1258096

Downie, A. W., Meiklejohn, G., St Vincent, L., Rao, A. R., Sundara Babu, B. V., and Kempe, C. H. (1965) Smallpox frequency and severity in relation to A, B and O blood groups. *Bulletin of the World Health Organization*, *33*(5), 623–625.

Epstein, A., A., Ottenberg, R. (1908) Simple method of performing serum reactions. *Proceedings of the NY Pathology Society*, 8, 117–123.

Fakhro, K. A., Staudt, M. R., Ramstetter, M. D., Robay, A., Malek, J. A., Badii, R., et al. (2016) The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Human Genome Variation*, *3*, 16016. http://doi.org/10.1038/hgv.2016.16

Feinstein, M., Ning, H., Kang, J., Bertoni, A., Carnethon, M., and Lloyd-Jones, D. M. (2012) Racial differences in risks for first cardiovascular events and noncardiovascular death: the Atherosclerosis Risk in Communities study, the Cardiovascular Health Study, and the Multi-Ethnic Study of Atherosclerosis. *Circulation*, *126*(1), 50–59. http://doi.org/10.1161/CIRCULATIONAHA.111.057232

Ferguson-Smith, M.A. Aitken, D.A. Turleau, C., De Grouchy, J. (1976) Localisation of the human ABO: Np-1: AK-1linkage group by regional assignment of AK-1–9q34. *Human Genetics*, 34, 35–43.

Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C., and Shendure, J. (2014) Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, *513*(7516), 120–123. http://doi.org/10.1038/nature13695

Florez, J. C., Price, A. L., Campbell, D., Riba, L., Parra, M. V., Yu, F., et al. (2009) Strong association of socioeconomic status with genetic ancestry in Latinos: implications for admixture studies of type 2 diabetes. *Diabetologia*, *52*(8), 1528–1536. http://doi.org/10.1007/s00125-009-1412-x

Fox, K., Johnsen, J. M., Coe, B., Frazar, C., Reiner, A., Eichler, E. E., and Nickerson, D. A. (2016) Analysis of exome sequencing datasets reveals structural variation in the coding region of ABO in individuals of African ancestry. *Transfusion*, (In press).

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., et al. (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American Journal of Human Genetics*, *91*(4), 597–607. http://doi.org/10.1016/j.ajhg.2012.08.005

Fry, A. E., Griffiths, M. J., Auburn, S., Diakite, M., Forton, J. T., Green, A., et al. (2008) Common variation in the ABO glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Human Molecular Genetics*, *17*(4), 567–576. http://doi.org/10.1093/hmg/ddm331

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., et al. (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, *493*(7431), 216–220. http://doi.org/10.1038/nature11690

Gagneux, P., and Varki, A. (1999) Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology*, *9*(8), 747–755. http://doi.org/10.1093/glycob/9.8.747

Georges, L., Seidenberg, V., Hummel, S., and Fehren-Schmitz, L. (2012) Molecular characterization of ABO blood group frequencies in pre-Columbian Peruvian highlanders. *American Journal of Physical Anthropology*, *149*(2), 242–249. http://doi.org/10.1002/ajpa.22115

Gini, C. (1968) Felix Bernstein. 1878-1956. *Genetics* (Vol. 60, pp. Suppl:22–3).

Giollo, M., Minervini, G., Scalzotto, M., Leonardi, E., Ferrari, C., and Tosatto, S. C. E. (2015) BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. *PloS One*, *10*(4), e0124579. http://doi.org/10.1371/journal.pone.0124579

Gloor, J. M., Lager, D. J., Moore, S. B., Pineda, A. A., Fidler, M. E., Larson, T. S., et al. (2003) ABO-incompatible kidney transplantation using both A2 and non-A2 living donors. *Transplantation*, *75*(7), 971–977. http://doi.org/10.1097/01.TP.0000058226.39732.32

Godber, M., Kopeć, A. C., Mourant, A. E., Teesdale, P., Tills, D., Weiner, J. S., et al. (1976) The blood groups, serum groups, red-cell isoenzymes and haemoglobins of the Sandawe and Nyaturu of Tanzania. *Annals of Human Biology*, *3*(5), 463–473.

Gordon, A. S., Fulton, R. S., Qin, X., Mardis, E. R., Nickerson, D. A., and Scherer, S. (2016) PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenetics and Genomics*, *26*(4), 161–168. http://doi.org/10.1097/FPC.0000000000000202

Halverson, M. S., and Bolnick, D. A. (2008) An ancient DNA test of a founder effect in Native American ABO blood group frequencies. *American Journal of Physical Anthropology*, *137*(3), 342–347. http://doi.org/10.1002/ajpa.20887

Holgersson, J., Clausen, H., Hakomori, S., Samuelsson, B. E., and Breimer, M. E. (1990) Blood group A glycolipid antigen expression in kidney, ureter, kidney artery, and kidney vein from a blood group A1Le(a-b+) human individual. Evidence for a novel blood group A heptaglycosylceramide based on a type 3 carbohydrate chain. *The Journal of Biological Chemistry*, *265*(34), 20790–20798.

Hosseini-Maaf, B., Irshaid, N. M., Hellberg, A., Wagner, T., Levene, C., Hustinx, H., et al. (2005) New and unusual O alleles at the ABO locus are implicated in unexpected blood group phenotypes. *Transfusion*, *45*(1), 70–81. http://doi.org/10.1111/j.1537-2995.2005.04195.x

Huang, P., Farkas, T., Marionneau, S., Zhong, W., Ruvoën-Clouet, N., Morrow, A. L., et al. (2003) Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: identification of 4 distinct strain-specific patterns. *The Journal of Infectious Diseases*, *188*(1), 19–31. http://doi.org/10.1086/375742

Hult, A. K., and Olsson, M. L. (2010) Many genetically defined ABO subgroups exhibit characteristic flow cytometric patterns. *Transfusion*, *50*(2), 308–323. http://doi.org/10.1111/j.1537-2995.2009.02398.x

Hult, A. K., Frame, T., Chesla, S., Henry, S., and Olsson, M. L. (2012) Flow cytometry evaluation of red blood cells mimicking naturally occurring ABO subgroups after modification with variable amounts of function-spacer-lipid A and B constructs. *Transfusion*, *52*(2), 247–251. http://doi.org/10.1111/j.1537-2995.2011.03268.x

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. http://doi.org/10.1038/nature06258

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., and Eichler, E. E. (2010) De novo rates and selection of large copy number variation. *Genome Research*, *20*(11), 1469–1481. http://doi.org/10.1101/gr.107680.110

James, A. B., Hillyer, C. D., and Shaz, B. H. (2012) Demographic differences in estimated blood donor eligibility prevalence in the United States. *Transfusion*, *52*(5), 1050–1061. http://doi.org/10.1111/j.1537-2995.2011.03416.x

Jeffreys, (1985) Individual-specific "fingerprints" of human DNA. *Nature*, 316 (1985), pp. 76–79.

Johnsen, J. M. (2015) Using red blood cell genomics in transfusion medicine. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, *2015*, 168–176. http://doi.org/10.1182/asheducation-2015.1.168

Johnsen, J. M., Auer, P. L., Morrison, A. C., Jiao, S., Wei, P., Haessler, J., et al. (2013a) Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: the NHLBI Exome Sequencing Project. *Blood*, *122*(4), 590–597. http://doi.org/10.1182/blood-2013-02-485094

Johnsen, J. M., Nickerson, D. A., and Reiner, A. P. (2013b) Massively parallel sequencing: the new frontier of hematologic genomics. *Blood*, *122*(19), 3268–3275. http://doi.org/10.1182/blood-2013-07-460287

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., and de Bakker, P. I. W. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics (Oxford, England)*, *24*(24), 2938–2939. http://doi.org/10.1093/bioinformatics/btn564

Kabat, E.A. (1956) Blood Group Substances. Their Chemistry and Immunochemistry. Academic Press, New York.

Khan, F., Khan, R. H., Sherwani, A., Mohmood, S., and Azfer, M. A. (2002) Lectins as markers for blood grouping. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, *8*(12), RA293–300.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. http://doi.org/10.1038/ng.2892

Koo, T. Y., and Yang, J. (2015) Current progress in ABO-incompatible kidney transplantation. *Kidney Research and Clinical Practice*, *34*(3), 170–179. http://doi.org/10.1016/j.krcp.2015.08.005

la Bastide, de, M., and McCombie, W. R. (2007) Assembling genomic DNA sequences with PHRAP. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [Et Al.]*, *Chapter 11*, Unit 11.4–11.4.15. http://doi.org/10.1002/0471250953.bi1104s17

Landsteiner, K. (1900) Zur Kenntniss der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zbl Bakt*, 27, 357–366.

Lane, W. J., Westhoff, C. M., Uy, J. M., Aguad, M., Smeland-Wagman, R., Kaufman, R. M., et al. (2016) Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion*, *56*(3), 743–754. http://doi.org/10.1111/trf.13416

Leffler, E. M., Gao, Z., Pfeifer, S., Ségurel, L., Auton, A., Venn, O., et al. (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science (New York, N.Y.)*, *339*(6127), 1578–1582. http://doi.org/10.1126/science.1234070

Linden, J. V., Wagner, K., Voytovich, A. E., and Sheehan, J. (2000) Transfusion errors in New York State: an analysis of 10 years' experience. *Transfusion*, *40*(10), 1207–1213.

Lindén, S., Mahdavi, J., Semino-Mora, C., Olsen, C., Carlstedt, I., Borén, T., and Dubois, A. (2008) Role of ABO secretor status in mucosal innate immunity and H. pylori infection. *PLoS Pathogens*, *4*(1), e2. http://doi.org/10.1371/journal.ppat.0040002

Llop, E., Henríquez, H., Moraga, M., Castro, M., and Rothhammer, F. (2006) Brief communication: Molecular characterization of O alleles at the ABO locus in Chilean Aymara and Huilliche Indians. *American Journal of Physical Anthropology*, *131*(4), 535–538. http://doi.org/10.1002/ajpa.20462

Londo, J., (2016) A time transect of exomes from a Native American population before and after European contact. (In progress.)

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science (New York, N.Y.)*, *335*(6070), 823–828. http://doi.org/10.1126/science.1215040

Mäkivuokko, H., Lahtinen, S. J., Wacklin, P., Tuovinen, E., Tenkanen, H., Nikkilä, J., et al. (2012) Association between the ABO blood group and the human intestinal microbiota composition. *BMC Microbiology*, *12*(1), 94. http://doi.org/10.1186/1471-2180-12-94

McBean, R. S., Hyland, C. A., and Flower, R. L. (2014) Approaches to determination of a full profile of blood group genotypes: single nucleotide variant mapping and massively parallel sequencing. *Computational and Structural Biotechnology Journal*, *11*(19), 147–151. http://doi.org/10.1016/j.csbj.2014.09.009

McCaughey, T., Liang, H. H., Chen, C., Fenwick, E., Rees, G., Wong, R. C. B., et al. (2016) An Interactive Multimedia Approach to Improving Informed Consent for Induced Pluripotent Stem Cell Research. *Cell Stem Cell*, *18*(3), 307–308. http://doi.org/10.1016/j.stem.2016.02.006

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. http://doi.org/10.1101/gr.107524.110

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science (New York, N.Y.)*, *338*(6104), 222–226. http://doi.org/10.1126/science.1224344

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., et al. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*(7332), 59–65. http://doi.org/10.1038/nature09708

Minikel, E. V., Vallabh, S. M., Lek, M., Estrada, K., Samocha, K. E., Sathirapongsasuti, J. F., et al. (2016) Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine*, *8*(322), 322ra9–322ra9. http://doi.org/10.1126/scitranslmed.aad5169

Morales, C. T., Muzquiz, L. I., Howlett, K., Azure, B., Bodnar, B., Finley, V., et al. (2016) Partnership with the Confederated Salish and Kootenai Tribes: Establishing an Advisory Committee for Pharmacogenetic Research. *Progress in Community Health Partnerships: Research, Education, and Action*, *10*(2), 173–183. http://doi.org/10.1353/cpr.2016.0035

Morgan, T. H. (1904) An analysis of the phenomena of organic 'polarity.' *Science (New York, N.Y.)*, *20*(518), 742–748. http://doi.org/10.1126/science.20.518.742

Morgan, W. T., and WATKINS, W. M. (2000) Unravelling the biochemical basis of blood group ABO and Lewis antigenic specificity. *Glycoconjugate journal* (Vol. 17, pp. 501–530)

Morgan, W. T. J. (1965) Blood group specific mucopolysaccharides. *Methods in Carbohydrate Chemistry*, 5,95–98.

Morgan, W. T. J. and Van Heyningen, R. (1944) The occurrence of A, B, and O blood group substances in pseudomucinous ovarian cyst fluids. *British Journal of Experimental Pathology*, 25, 5–15.

Moulds, J. M., and Moulds, J. J. (2000) Blood group associations with parasites, bacteria, and viruses. *Transfusion Medicine Reviews*, *14*(4), 302–311. http://doi.org/10.1053/tmrv.2000.16227

Moulds, J. M., Nowicki, S., Moulds, J. J., and Nowicki, B. J. (1996) Human blood groups: incidental receptors for viruses and bacteria. *Transfusion*, *36*(4), 362–374. http://doi.org/10.1046/j.1537-2995.1996.36496226154.x

Mountain, J. L., and CAVALLI-SFORZA, L. L. (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(14), 6515–6519.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, N.Y.)*, *337*(6090), 100–104. http://doi.org/10.1126/science.1217876

Nilsson, C. L. (2007) Lectins: analytical technologies (1st ed.). Elsevier. ISBN 9780444530776 Boston, Massachusetts, USA

Noderer, W. L., Flockhart, R. J., Bhaduri, A., Diaz de Arce, A. J., Zhang, J., Khavari, P. A., and Wang, C. L. (2014) Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Molecular Systems Biology*, *10*(8), 748–748. http://doi.org/10.15252/msb.20145136

Nydegger, U., Mohacsi, P., Koestner, S., Kappeler, A., Schaffner, T., and Carrel, T. (2005) ABO histo-blood group system-incompatible allografting. *International Immunopharmacology*, *5*(1), 147–153. http://doi.org/10.1016/j.intimp.2004. 09.020

O'Roak, B. J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I. G., et al. (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science, 338(6114), 1619-1622. http://doi.org/10.1126/science. 1227764

Oh, S. S., Galanter, J., Thakur, N., Pino-Yanes, M., Barcelo, N. E., White, M. J., et al. (2015) Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. *PLoS Medicine*, *12*(12), e1001918. http://doi.org/10.1371/journal.pmed.1001918

Olsson, M. L., and Clausen, H. (2008) Modifying the red cell surface: towards an ABO-universal blood supply. *British Journal of Haematology*, *140*(1), 3–12. http://doi.org/10.1111/j.1365-2141.2007.06839.x

Oriol, R., Samuelsson, B. E., and Messeter, L. (1990) ABO antibodies--serological behaviour and immuno-chemical characterization. *Journal of Immunogenetics*, *17*(4-5), 279–299.

Paré, G., Chasman, D. I., Kellogg, M., Zee, R. Y. L., Rifai, N., Badola, S., et al. (2008) Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genetics*, *4*(7), e1000118. http://doi.org/10.1371/journal.pgen.1000118

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., et al. (1998) Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*, *63*(6), 1839–1851. http://doi.org/10.1086/302148

Patenaude, S. I., Seto, N. O. L., Borisova, S. N., Szpacenko, A., Marcus, S. L., Palcic, M. M., and Evans, S. V. (2002) The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nature Structural Biology*, *9*(9), 685–690. http://doi.org/10.1038/nsb832

Patnaik, S. K., Helmberg, W., and Blumenfeld, O. O. (2014) BGMUT Database of Allelic Variants of Genes Encoding Human Blood Group Antigens. *Transfusion Medicine and Hemotherapy : Offizielles Organ Der Deutschen Gesellschaft Fur Transfusionsmedizin Und Immunhamatologie*, *41*(5), 346–351. http://doi.org/10.1159/000366108

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. http://doi.org/10.1038/nature12886

Putkonen, T. (1930) Über gruppenspezifischen eigenschaften verschiedener körper flussigkeiten. *Acta Society Med. Fenn 'Duodecim' A*, 14 (2).

Racaniello, V. R., and Baltimore, D. (1981) Cloned poliovirus complementary DNA is infectious in mammalian cells. *Science (New York, N.Y.)*, *214*(4523), 916–919.

Ravn, V., and Dabelsteen, E. (2000) Tissue distribution of histo-blood group antigens. *APMIS : Acta Pathologica, Microbiologica, Et Immunologica Scandinavica*, *108*(1), 1–28.

Reardon, S. (2015, July 23) US tailored-medicine project aims for ethnic balance. *Nature*, pp. 391–392. http://doi.org/10.1038/523391a

Rieder, M. J., Reiner, A. P., Gage, B. F., Nickerson, D. A., Eby, C. S., McLeod, H. L., et al. (2005) Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *The New England Journal of Medicine*, *352*(22), 2285–2293. http://doi.org/10.1056/NEJMoa044503

Rochant H. (1976) Abnormal Distribution of Erythrocytes A_1 Antigens in Preleukemia as Demonstrated by an Immunofluorescence Technique. Hemopoietic Dysplasias (Preleukemic States) Section 2.2. pp 237-255

Romano, E. L., Soyano, A., Montaño, R. F., Ratcliffe, M., Olson, M., Suarez, G., et al. (1994) Treatment of ABO hemolytic disease with synthetic blood group trisaccharides. *Vox Sanguinis*, *66*(3), 194–199.

Roubinet, F., Janvier, D., and Blancher, A. (2002) A novel cis AB allele derived from a B allele through a single point mutation. *Transfusion*, *42*(2), 239–246.

Saitou, N., and Yamamoto, F. (1997) Evolution of primate ABO blood group genes and their homologous genes. *Molecular Biology and Evolution*, *14*(4), 399–411.

Sano, R., Kuboya, E., Nakajima, T., Takahashi, Y., Takahashi, K., Kubo, R., et al. (2015) A 3.0-kb deletion including an erythroid cell-specific regulatory element in intron 1 of the ABO blood group gene in an individual with the Bm phenotype. *Vox Sanguinis*, *108*(3), 310–313. http://doi.org/10.1111/vox.12216

Scheet, P., and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*(4), 629–644. http://doi.org/10.1086/502802

Seffens, W., Evans, C., Minority Health-GRID Network, and Taylor, H. (2015) Machine Learning Data Imputation and Classification in a Multicohort Hypertension Clinical Study. *Bioinformatics and Biology Insights*, *9*(Suppl 3), 43–54. http://doi.org/10.4137/BBI.S29473

Ségurel, L., Thompson, E. E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S. W., et al. (2012) The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(45), 18493–18498. http://doi.org/10.1073/pnas.1210603109

Shaz, B. H., James, A. B., Hillyer, K. L., Schreiber, G. B., and Hillyer, C. D. (2010) Demographic variations in blood donor deferrals in a major metropolitan area. *Transfusion*, *50*(4), 881–887. http://doi.org/10.1111/j.1537-2995.2009.02501.x

Shin, M., and Kim, S.-J. (2011) ABO Incompatible Kidney Transplantation-Current Status and Uncertainties. *Journal of Transplantation*, *2011*(6), 970421–11. http://doi.org/10.1155/2011/970421

Smith, N. L., Chen, M.-H., Dehghan, A., Strachan, D. P., Basu, S., Soranzo, N., et al. (2010) Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation*, *121*(12), 1382–1392. http://doi.org/10.1161/CIRCULATIONAHA.109.869156

Snyder, M. W., Adey, A., Kitzman, J. O., and Shendure, J. (2015) Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews. Genetics*, *16*(6), 344–358. http://doi.org/10.1038/nrg3903

Sorensen, J. B., Grant, W. J., Belnap, L. P., Stinson, J., and Fuller, T. C. (2001) Transplantation of ABO group A2 kidneys from living donors into group O and B recipients. *American Journal of Transplantation : Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, *1*(3), 296–299.

Sousa, N. C., Anicchino-Bizzacchi, J. M., Locatelli, M. F., Castro, V., and Barjas-Castro, M. L. (2007) The relationship between ABO groups and subgroups, factor VIII and von Willebrand factor. *Haematologica*, *92*(2), 236–239.

Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., et al. (2015) Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*, *200*(2), 413–422. http://doi.org/10.1534/genetics.115.175802

Storry, J. R., and Olsson, M. L. (2004) Genetic basis of blood group diversity. *British Journal of Haematology*, *126*(6), 759–771. http://doi.org/10.1111/j.1365-2141.2004.05065.x

Storry, J. R., Johannesson, J. S., Poole, J., Strindberg, J., Rodrigues, M. J., Yahalom, V., et al. (2006) Identification of six new alleles at the FUT1 and FUT2 loci in ethnically diverse individuals with Bombay and Para-Bombay phenotypes. *Transfusion*, *46*(12), 2149–2155. http://doi.org/10.1111/j.1537-2995.2006.01045.x

Storry, J. R., Olsson, M. L. (2009) The ABO blood group system revisited: a review and update. Immunohematology, 25(2):48-59.

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., et al. (2013) Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*, *23*(9), 1373–1382. http://doi.org/10.1101/gr.158543.113

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. http://doi.org/10.1038/nature15394

Sukernik, R. I., Karaphet, T. M., and Osipova, L. P. (1978) Distribution of blood groups, serum markers and red cell enzymes in two human populations from Northern Siberia. *Human Heredity*, *28*(5), 321–327.

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., and Salim, A. (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics (Oxford, England)*, *28*(21), 2711–2718. http://doi.org/10.1093/bioinformatics/bts535

Tishkoff, S. (2015) GENETICS: Strength in small numbers. *Science (New York, N.Y.)*, *349*(6254), 1282–1283. http://doi.org/10.1126/science.aad0584

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., et al. (2009) The genetic structure and history of Africans and African Americans. *Science (New York, N.Y.)*, *324*(5930), 1035–1044. http://doi.org/10.1126/science.1172257

Trégouët, D.-A., Heath, S., Saut, N., Biron-Andreani, C., Schved, J.-F., Pernod, G., et al. (2009) Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood*, *113*(21), 5298–5303. http://doi.org/10.1182/blood-2008-11-190389

Vernot, B., Tucci, S., Kelso, J., Schraiber, J. G., Wolf, A. B., Gittelman, R. M., et al. (2016) Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science (New York, N.Y.)*, *352*(6282), 235–239. http://doi.org/10.1126/science.aad9416

Villanea, F. A., Bolnick, D. A., Monroe, C., Worl, R., Cambra, R., Leventhal, A., and Kemp, B. M. (2013) Brief communication: Evolution of a specific O allele (O1vG542A) supports unique ancestry of Native Americans. *American Journal of Physical Anthropology*, *151*(4), 649–657. http://doi.org/10.1002/ajpa.22292

Vogt, N. (2016) Synthetic biology: Customizing cell-cell communication. *Nature Methods*, *13*(4), 285.

Von Decastello, A. and Sturli, A. (1902) Über die isoagglutinine im serum gesunder und kranker Menschen.*Munchen Medicine Wchnschr*, 49, 1090–1095.

Von Dungern, E. and Hirszfeld, L. (1910) Über Vererbung gruppenspezifischer Strukturen des Blutes. *Zeitscrift für Immunforsch*, 6, 284–292.

Wang, J., Fan, H. C., Behr, B., and Quake, S. R. (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, *150*(2), 402–412. http://doi.org/10.1016/j.cell.2012.06.030

Watkins, W. M. (1966) Blood group substances. *Science*, 152, 172–181.

Watkins, W. M. (1980) Biochemistry and genetics of the ABO, Lewis and P blood group systems. *Advances in Human Genetics*, 10, 1–136.

Wessel, J., Chu, A. Y., Willems, S. M., Wang, S., Yaghootkar, H., Brody, J. A., et al. (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature Communications*, *6*, 5897. http://doi.org/10.1038/ncomms6897

West, L. J., Karamlou, T., Dipchand, A. I., Pollock-BarZiv, S. M., Coles, J. G., and McCrindle, B. W. (2006) Impact on outcomes after listing and transplantation, of a strategy to accept ABO blood group-incompatible donor hearts for neonates and infants. *The Journal of Thoracic and Cardiovascular Surgery*, *131*(2), 455–461. http://doi.org/10.1016/j.jtcvs.2005.09.048

West, L. J., Pollock-Barziv, S. M., Dipchand, A. I., Lee, K. J., Cardella, C. J., Benson, L. N., et al. (2001) ABO-incompatible heart transplantation in infants. *The New England Journal of Medicine*, *344*(11), 793–800. http://doi.org/10.1056/NEJM200103153441102

Wiggins, K. L., Smith, N. L., Glazer, N. L., Rosendaal, F. R., Heckbert, S. R., Psaty, B. M., et al. (2009) ABO genotype and risk of thrombotic events and hemorrhagic stroke. *Journal of Thrombosis and Haemostasis : JTH*, *7*(2), 263–269. http://doi.org/10.1111/j.1538-7836.2008.03243.x

Wolofsky, K. T., Ayi, K., Branch, D. R., Hult, A. K., Olsson, M. L., Liles, W. C., et al. (2012) ABO blood groups influence macrophage-mediated phagocytosis of Plasmodium falciparum-infected erythrocytes. *PLoS Pathogens*, *8*(10), e1002942. http://doi.org/10.1371/journal.ppat.1002942

Wolpin, B. M., Kraft, P., Xu, M., Steplowski, E., Olsson, M. L., Arslan, A. A., et al. (2010) Variant ABO blood group alleles, secretor status, and risk of pancreatic cancer: results from the pancreatic cancer cohort consortium. *Cancer Epidemiology, Biomarkers and Prevention : a Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *19*(12), 3140–3149. http://doi.org/10.1158/1055-9965.EPI-10-0751

Woodahl, E. L., Lesko, L. J., Hopkins, S., Robinson, R. F., Thummel, K. E., and Burke, W. (2014) Pharmacogenetic research in partnership with American Indian and Alaska Native communities. *Pharmacogenomics*, *15*(9), 1235–1241. http://doi.org/10.2217/pgs.14.91

Yamamoto, F. (1992) [Molecular biology of ABO genes]. *Tanpakushitsu Kakusan Koso. Protein, Nucleic Acid, Enzyme*, *37*(11 Suppl), 1696–1700.

Yamamoto F. (2012) ABO in the Context of Blood Transfusion and Beyond. Blood Transfusion in Clinical Practice, book edited by Puneet Kaur Kochhar, ISBN 978-953-51-0343-1, Published: March. DOI: 10.5772/35812

Yamamoto, F., and Hakomori, S. (1990) Sugar-nucleotide donor specificity of histo-blood group A and B transferases is based on amino acid substitutions. *The Journal of Biological Chemistry*, *265*(31), 19257–19262.

Yamamoto, F., Cid, E., Yamamoto, M., and Blancher, A. (2012) ABO research in the modern era of genomics. *Transfusion Medicine Reviews*, *26*(2), 103–118. http://doi.org/10.1016/j.tmrv.2011.08.002

Yamamoto, F., Cid, E., Yamamoto, M., Saitou, N., Bertranpetit, J., and Blancher, A. (2014) An integrative evolution theory of histo-blood group ABO and related genes. *Scientific Reports*, *4*, 6601. http://doi.org/10.1038/srep06601

Yamamoto, F., Clausen, H., White, T., Marken, J., and Hakomori, S. (1990) Molecular genetic basis of the histo-blood group ABO system. *Nature*, *345*(6272), 229–233. http://doi.org/10.1038/345229a0

Yamamoto, F., McNeill, P. D., and Hakomori, S. (1992) Human histo-blood group A2 transferase coded by A2 allele, one of the A subtypes, is characterized by a single base deletion in the coding sequence, which results in an additional domain at the carboxyl terminal. *Biochemical and Biophysical Research Communications*, *187*(1), 366–374.

Yamamoto, F., McNeill, P. D., and Hakomori, S. (1995) Genomic organization of human histo-blood group ABO genes. *Glycobiology*, *5*(1), 51–58.

Yazdanbakhsh, K., Ware, R. E., and Noizat-Pirenne, F. (2012) Red blood cell alloimmunization in sickle cell disease: pathophysiology, risk factors, and transfusion management. *Blood*, *120*(3), 528–537. http://doi.org/10.1182/blood-2011-11-327361

Yazer, M. H., Hosseini-Maaf, B., and Olsson, M. L. (2008a) Blood grouping discrepancies between ABO genotype and phenotype caused by O alleles. *Current Opinion in Hematology*, *15*(6), 618–624. http://doi.org/10.1097/MOH. 0b013e3283127062

Yazer, M. H., Hult, A. K., Hellberg, A., Hosseini-Maaf, B., Palcic, M. M., and Olsson, M. L. (2008b) Investigation into A antigen expression on O2 heterozygous group O-labeled red blood cell units. *Transfusion*, *48*(8), 1650–1657. http://doi.org/10.1111/j.1537-2995.2008.01732.x

Yen J., (2016) A variant by any name: quantifying annotation discordance across tools and clinical databases. Biorxiv. doi: http://dx.doi.org/10.1101/054023

Yip, S. P. (2002) Sequence variation at the human ABO locus. *Annals of Human Genetics*, *66* (Pt 1), 1–27. http://doi.org/10.1017/S0003480001008995

Zabaneh, D., Gaunt, T. R., Kumari, M., Drenos, F., Shah, S., Berry, D., et al. (2011) Genetic variants associated with Von Willebrand factor levels in healthy men and women identified using the HumanCVD BeadChip. *Annals of Human Genetics*, *75*(4), 456–467. http://doi.org/10.1111/j.1469-1809.2011.00654.x

Zhang, H., Mooney, C. J., and Reilly, M. P. (2012) ABO Blood Groups and Cardiovascular Diseases. *International Journal of Vascular Medicine*, *2012*(3), 641917–11. http://doi.org/10.1155/2012/641917