

https://dspace.org/

Antibody Genetic Diversity within an African population

Marsden, Alaine Athenaïs

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/49988
Marsden, Alaine Athenaïs.(2025). The role of design houses [PhD thesis, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/49988

Downloaded from a DSpace repository. https://dspace.org/

Antibody Genetic Diversity within an African population

Alaine Athenaïs Marsden (706359)

A Thesis Submitted to the Faculty of Health Sciences University of the Witwatersrand

In Fulfillment of the Requirements

For Doctor of Philosophy in Virology

March 2025

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

Candidate Declaration

I, Alaine (Joshua) Marsden, declare that this Thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy in Virology at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

_____

On this, the _____day of March 2025 in Johannesburg, South Africa

Abstract

Antibodies, also referred to as immunoglobulins, are a critical part of the adaptive immune response. They are produced through somatic recombination of immunoglobulin (IG) genes from genetically complex IG loci to form two chains - the heavy chain (IGH locus) and light chain (IGK and IGL loci). These highly variable IG loci contain extensive polymorphic diversity and structural variation; where insertion, deletion, and duplication events create copy number variation (CNV) in several IG genes. There is increasing evidence that genetic variation in antibody genes affects their expression, in addition to their affinity to bind antigens, and engage the broader immune system. In this thesis, we utilise transcriptomic and genomic techniques to characterise IG genetic diversity in a South African population. We examined the genes that encode the antigen binding site (variable genes: IGHV, IGKV, and IGLV), as well as those that interact with the broader immune system (constant genes: IGHG and IGHA). Chapter 1 examines the heavy chain variable (IGHV) genes, uncovering 53 novel alleles, including some with novel single nucleotide variants (SNVs) in the region of the gene that regulates antibody secretion. In addition to novel diversity, gene duplication was common in this African population, occurring in 18 genes of which ten are newly described. Furthermore, we provide evidence that duplication can enhance the expression of particular IGHV genes. In Chapter 2, we sequenced the IGHA and IGHG genes, that code for the IgA and IgG isotypes respectively. Genetic diversity in IGHA is poorly studied and here we report the addition of 26 novel alleles. While IGHG is more well-studied, this study provides an additional 29 novel genetic variants in this gene. The allelic variation uncovered for both of these genes resulted in protein variation that can affect how antibodies bind Fc receptors, enhancing or inhibiting effector functions, or altering the structural aspects of antibodies such as dimerisation. Chapter 3 examined the expressed light chain lambda and kappa (IGKV and IGLV) repertoires of three different populations (South Africa, Norway, and the United Kingdom). Despite significant genetic differences between these groups, the same limited set of light chain genes dominate the expressed repertoires of each group, with only five IGKV genes representing 50% of the expressed repertoire and only two IGLV genes accounting for between 20-50% of the repertoire. Overall, we report significant amounts of undiscovered genetic diversity in South African IG genes, uncovering 118 novel alleles across different IG genes. This variation can strongly affect the functional antibody repertoire. Understanding these dynamics will have implications for broader immunology and vaccinology, especially with the advent of germline targeting vaccine strategies for HIV and other pathogens. The work in this thesis serves as the foundation for further functional study of genetic IG variation.

Acknowledgements and Dedication

This thesis is dedicated to my parents, whose continuous support provides the bedrock for all that I do.

“Hear, my son, your father's instruction, and forsake not your mother's teaching, for they are a graceful garland for your head and pendants for your neck.”

Proverbs 1:8-9

I would firstly like to acknowledge my supervisors for everything that they have done for me over the past few difficult years. All of the advice, the many drafts, the endless meetings, the tireless pushing - it all makes me a better scientist and person.

I would also like to acknowledge various members of the AIRR-C community for the conversations, seminars, emails and input over the years in which this research came together: Dr Corey Watson, Dr Ivanna Mikocziova, Prof. Gunilla Karlsson Hedestam, Prof. Mats Ohlin, Dr William Gibson, Dr Oscar Rodriguez, amongst others.

Plagiarism Declaration

I, Alaine Marsden (Student number: 706359) am a student registered for the degree of Doctor of Philosophy in the academic year 2024.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work
- except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: _____ Date: _____

Table of Contents

Candidate Declaration ..... **ii**
Abstract ..... **iii**
Acknowledgements and Dedication ..... **iv**
Plagiarism Declaration ..... **v**
Table of Contents ..... **vi**
List of Figures ..... **ix**
List of Tables ..... **x**
Research Outputs ..... **xi**
Acknowledgement of Funding ..... **xi**

Introduction 1

Basic Antibody Biology **1**
Structure and Function of Antibodies 1
Generation of the Antigen Binding Region 3
Antibody Genetics and Human Diversity **5**
Diversity in the Immunoglobulin (IG) Genetic Loci 5
Immunogenetic sequencing techniques 7
IG Gene Nomenclature 7
The Antibody Response in HIV **9**
Chapter Overview **10**
References **11**

Samples and Ethics 16

Ethics Declaration 16
CAPRISA Participant Information 16

Chapter 1 (Published Journal Article):

Novel polymorphic and copy number diversity in the antibody IGH locus of South African Individuals **19**

Chapter 2:
Identification of novel IgA and IgG antibody alleles in a South African population.....32

Abstract.....32
Introduction.....33
Methods.....35
Participant samples and ethics.....35
Germline IGHA library preparation and sequencing.....35
Data analysis.....37
Results.....39
Overview of IGHA and IGHG sequencing.....39
IGHA1 nucleotide and amino acid variation.....40
IGHG nucleotide and amino acid variation.....44
Discussion.....50
References.....54

Chapter 3:
Meta-analysis of IGLV and IGKV repertoires from three distinct populations.....57

Abstract.....57
Introduction.....58
Methods.....61
Participant samples and ethics.....61
Preprocessing of raw sequences.....61
Analysis of IGLV and IGKV sequences.....62
Results.....63
IGKV and IGLV allelic diversity.....63
IGKV and IGLV gene usage biases between different populations.....65
Expression of genes associated with public anti-HIV antibodies for germline targeting immunogen design.....70
Discussion.....73
References.....76

Conclusion.....79

References.....84

Thesis References……88

Addenda…..……97

Addendum A: Ethics Clearance Certificates.....97
Wits HREC Certificate for this study 97
Wits HREC Certificate for CAPRISA cohorts.....98
UKZN BREC Certificate for CAPRISA cohorts 99
Addendum B: Supplementary Data from Chapter 1 100
Supplementary Table 1: Primers for germline NGS.....100
Supplementary Table 2: IGHV Alleles Detected 102
Supplementary Table 3: Identity of Novel IGHV Alleles.....104
Supplementary Table 4: IGHV alleles with matching V regions showing mutations in other regions 108
Supplementary Table 5: IGHV Leader Sequences.....110
Supplementary Table 6: Alleles of duplicated IGHV genes from genomic data 113
Supplementary Table 7: RabHit output table 119
Addendum C: Plagiarism Report.....120

List of Figures

Figures in Introduction and General Literature Review

Figure 1: Schematic representation of an antibody. 2
Figure 2: Schematic representation of VDJ recombination....3
Figure 3: A representation of structural variation in the IGH locus demonstrating insertion and deletion of IGHV genes between different haplotypes....6
Figure 4: Breakdown of IG gene naming conventions. 8

Figures in Chapter 2

Figure 1: Description of IGHA and IGHG germline sequencing....39
Figure 2: Distribution of nucleotide diversity in IGHA....40
Figure 3: Novel allotypic variation in IGHA1 42
Figure 4: Nucleotide diversity in IGHG1 and IGHG2 genes....45
Figure 5: Nucleotide diversity in IGHG3 and IGHG4 genes....46
Figure 6: Allotypes in IGHG 48

Figures in Chapter 3

Figure 1: Distribution of IGKV and IGLV alleles. 64
Figure 2: IGKV gene usage biases....66
Figure 3: IGKV alleles for prominently expressed genes. 66
Figure 4: IGLV gene usage biases. 69
Figure 5: IGLV alleles for prominently expressed genes. 70

List of Tables

Table in Samples and Ethics

CAPRISA Participant Information 16

Tables in Chapter 2

Table 1: Alleles associated with IGHA1 allotypes 43

Table 2: Alleles associated with IGHG allotypes 49

Table 3: Summary of mutations in IGHG and IGHA sequences from this study and their potential effects 52

Tables in Chapter 3

Table 1: HIV bNAb lineages prioritised for germline targeting immunogens 72

Research Outputs

Published Papers

Marsden, A.A., Corcoran, M., Hedestam, G.K. et al. Novel polymorphic and copy number diversity in the antibody IGH locus of South African individuals. Immunogenetics 77, 6 (2025). https://doi.org/10.1007/s00251-024-01363-7

Oral Presentations

Scripps Consortium For HIV/AIDs Vaccine Development (CHAVD) Principal Investigators Meeting - 17 October 2022

ADVANCE (Accelerate the Development of Vaccines and New Technologies to Combat the AIDS Epidemic) Repertoire Working Group Seminar - 13 December 2024

Poster Presentations

International Congress for Human Genetics 2023 (ICHG2023) - February 2023

Collaboration for AIDS Vaccine Discovery (CAVD) Africa Regional Meeting - February 2023

Acknowledgement of Funding

Funding was provided through the H3Africa consortium: NIAID (U01AI136677) and the SA Medical Research Council SHIP program. A student bursary was provided by the Poliomyelitis Research Foundation: 22/81

Introduction and General Literature Review

A broad overview of antibody biology and genetics

Basic Antibody Biology

Structure and Function of Antibodies

Antibodies, or immunoglobulins, are a key part of the adaptive immune response (1). These Y-shaped proteins are produced by B-cells, a population of mononuclear lymphocytes. Antibodies play a large role in recognising pathogens and signalling the immune system to destroy the invasive entity (2). They are produced in response to an infection and adapt to evolving threats such as viruses. The production of antibodies has been demonstrated as a correlate for protection when assessing the efficacy of vaccines (3,4).

An antibody is formed by two identical pairs of heavy and light chains, (Figure 1) (5). There are two different portions in an antibody, the fragment antigen binding (Fab) portion and the fragment crystallizable (Fc) portion. It is the Fab portion of the antibody that binds to pathogen surface molecules and other antigens, while the Fc portion interfaces with the broader immune system to elicit effector functions (6). The Fc of the antibody also defines the isotype or class. There are five main isotypes: IgM, IgA, IgG, IgE and IgD (7). The IgM isotype is the first class of antibody that is produced by an antigen naive B-cell (8). Mucosal immunity is largely mediated by IgA antibodies (9). The most protective and most abundant class of antibody is IgG, which is predominantly found circulating in the blood (10). IgE originally evolved to combat parasitic worms, however in modern humans it has been linked to allergies and diseases like asthma (11). The last isotype, IgD, is co-expressed with IgM and aids in the development of antigen naive B-cells (12).

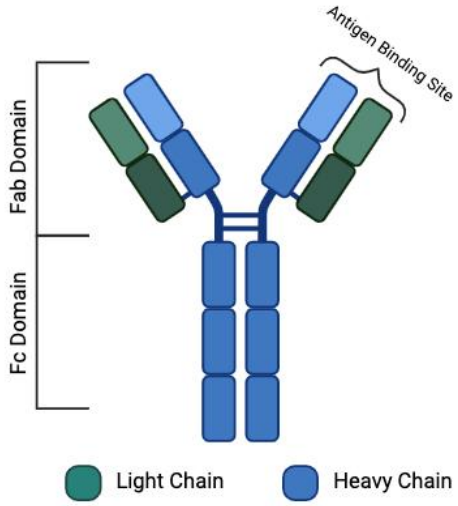


Figure 1: Schematic representation of an antibody. The heavy chain is shown in blue and the light chain is shown in green.

Antibodies perform a variety of functions within the adaptive immune response. Firstly, antibodies are able to bind surface proteins on pathogens that are necessary for cellular entry, thus physically inhibiting that pathogen from infecting the host cell (13). This is termed neutralisation. Antibodies are also able to elicit effector function(s) through signalling to other cells in the immune system (14). Antibody dependent cellular cytotoxicity (ADCC) involves recruiting other immune cells, such as natural killer cells, which target infected host cells with directed release of cytotoxic granules that contain perforin and granzymes (15). This process causes lysis of the target pathogen. The target pathogen can also be engulfed by an immune cell via antibody dependent cellular phagocytosis (ADCP), the pathogen is then destroyed within a lysosome inside the phagocytic cell (16). An antibody is able to activate complement through binding of C1q, which causes a complement cascade that results in the pathogen being lysed by membrane attack complexes (17).

Effector functions are enabled through binding of immune receptors to the Fc portion of the antibody via Fc receptors (FcRs). Different isotypes of antibodies have variable affinity for the FcRs present on various cell types (18). This causes specific antibody isotypes to be more effective at eliciting certain responses over others. For example, Fcγ RIIIA, which is expressed by NK cells, is mostly strongly bound by IgG3 leading to stronger ADCC responses (19). Recent studies have also indicated that allelic diversity of constant genes can also affect FcR binding (17,20,21). This is discussed more extensively in Chapter 2.

Generation of the Antigen Binding Region

The expressed antibody repertoire has the capacity to recognise and adapt to a wide number of potential threats to the human body (1). The initial source of this capability arises from V(D)J recombination (Figure 2). This is a somatic reorganisation event that occurs during the early development of a B-cell where one variable (V) gene, one diversity (D) gene and one joining (J) gene are joined together to form the antigen binding part of an antibody (22). This process is facilitated by recombinase activating gene (Rag) enzymes that recognise recombination signal sequences in IG genes and combine them together (23). The IG genes are selected semi-randomly from a larger pool of V, D, and J genes, creating a broad array of combinatorial diversity in the naive antibody repertoire. The rate at which certain VDJ combine with each other is not entirely random as chromosomal location and chromatin conformation have been shown to impact the way in which somatic recombination occurs along the loci (24–26).

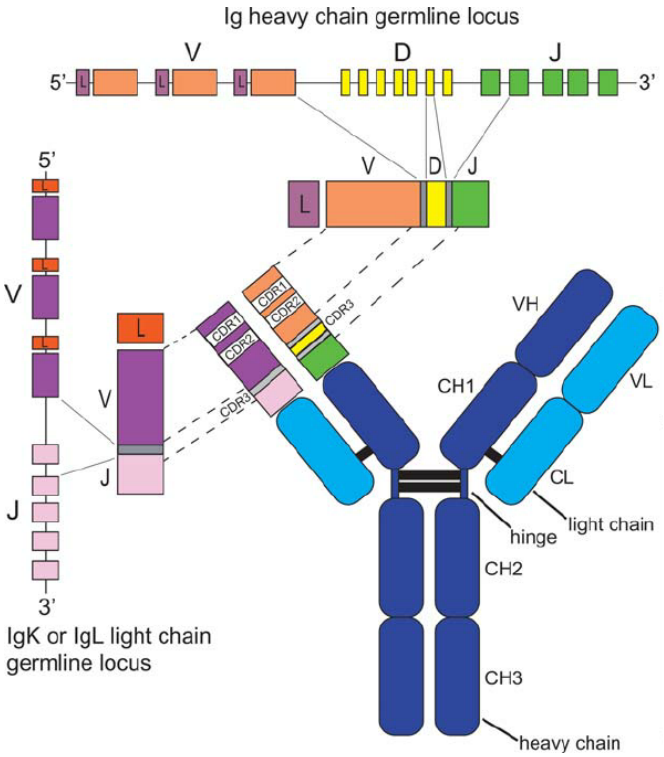


Figure 2: Schematic representation of VDJ recombination. Adapted from (27).

The binding of an antibody to a pathogen is further refined through a process called affinity maturation. When a naive IgM binds to a potential target, it undergoes clonal expansion after stimulation by regulatory T-cells (8). During this process, an enzyme called Activation-induced cytidine deaminase (AID) causes somatic hypermutation of the antibody sequence, altering

nucleotide positions in the antibody sequence (28). Enzymatically, AID deaminates cytosine, changing the nucleotide into thymine. This DNA damage is repaired by error prone DNA repair enzymes leading to the creation of a variety of different mutations within the antibody sequence. These mutations are concentrated in certain “hot spots” within the antibody sequence, primarily the complementarity determining regions (CDRs) which play the largest role in antigen binding (29). The affinity of the new B-cell will be assessed by the immune system, where antibodies with high affinity for the target pathogen undergo further clonal expansion and antibodies with poor affinity are discarded (2). Another process that occurs during clonal expansion is class switch recombination, which changes the antibody isotype by swapping the constant gene of the heavy chain, which encodes the Fc portion of an antibody (30). These processes alter and refine antibodies to allow the immune system to respond and adapt to evolving pressure from invasive pathogens.

Antibody Genetics and Human Diversity

Diversity in the Immunoglobulin (IG) Genetic Loci

The genes which encode antibodies (IG genes), are contained within different genetic loci that occur across the human genome. The immunoglobulin heavy chain (IGH) locus is positioned on the telomeric extremity of the q arm of chromosome 14 (31). All of the genes which make up the heavy chain are found in this IGH locus (5). This includes the variable (IGHV), diversity (IGHD) and joining (IGHJ) genes which form together to make the antigen binding part of an antibody. The constant genes which encode the Fc region are also contained in this locus (IGHM, IGHD, IGHG, IGHA and IGHE) and correspond to their relevant isotypes.

The genes which form the light chain are found in two different genetic loci, leading to the heavy chain being paired with a light chain that is produced by either locus (32). The immunoglobulin kappa (IGK) locus is found on the p arm of chromosome 2 and the immunoglobulin lambda (IGL) locus is found on the q arm of chromosome 22. These loci contain variable (IGKV/IGLV) and joining (IGKJ/IGLJ) genes. There are constant genes (IGKC / IGLC) within the light chain loci but they do not contribute to the Fc region of the antibody.

The IG loci demonstrate incredible genetic diversity (31,32). There are two main forms of this genetic diversity, nucleotide polymorphism and structural variation. The widespread polymorphic diversity in the IG loci produces many alleles for IG genes, especially the variable V genes (33). Structural variation within the IG loci results from large scale insertion, duplication, inversion and deletion of genomic segments (Figure 3) (34). These can also involve more complex variants that are not resolved through a single or pair of breakpoints. Structural variants often involve several IG genes, creating alterations in copy number variation for up to four genes at a time. The impact of genetic diversity, both through polymorphism and structural variation, has become a key interest in the field of immunogenetics. Specific examples and studies for the genes studied in this thesis are in the introductions for each chapter. These are important considerations as host genetics may influence the way in which a particular individual responds to either infection, thus impacting disease progression, or vaccination, affecting the efficacy of the vaccine response.

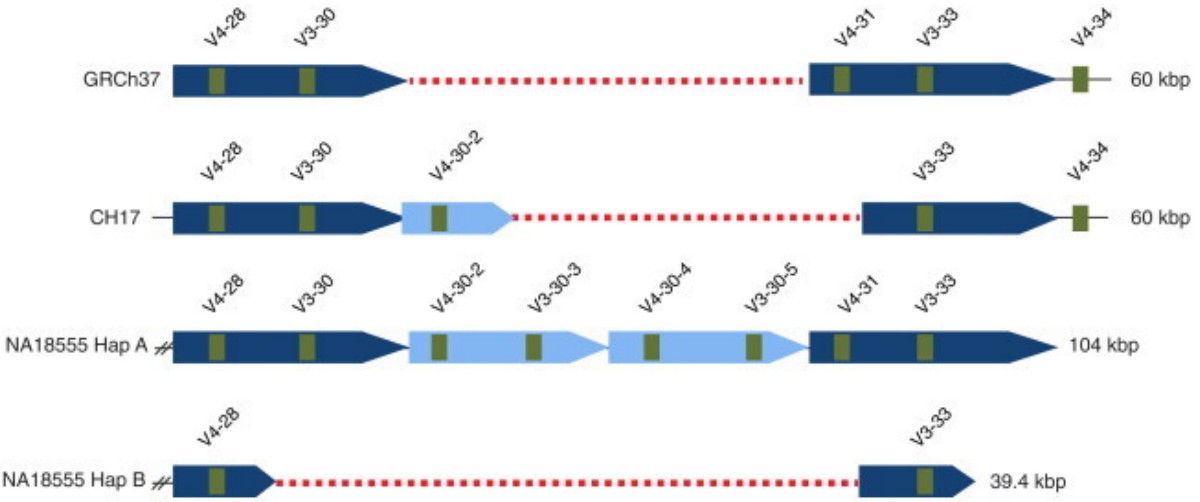


Figure 3: A representation of structural variation in the IGH locus demonstrating insertion and deletion of IGHV genes between different haplotypes. Inserted genes are shown in light blue and deletions are shown as dashed red lines. Adapted from (35).

Over the past decade there has been increasing momentum in the immunogenetic community to understand human genetic diversity. These studies have examined a wide variety of IG genes. A large portion of studies that aim to characterise novel IG gene diversity focus on IGHV (35–41). There are fewer studies that examine light chain variable gene diversity (IGLV and IGKV) (42–45). Lastly, studies that sequence the constant genes, which encode the Fc, are rare and predominantly focus on IGHG which codes for the IgG isotype (46,47). These studies have bolstered public databases with many new alleles for IG genes as sequencing techniques have allowed for more individuals to be sequenced across more populations. Novel structural variants (large scale duplications, deletions and insertions) have also been uncovered through long read sequencing methods, adding to the complexity of genetic diversity in the IG loci (35,39,42,43,45).

African diversity in the IG loci is less well understood. Prior studies have either utilised Yoruban samples from the 1000 genomes project or studied South African populations (35,38,39,42,43,45,47). The population diversity across the continent of Africa is incredibly complex and demonstrates higher degrees of genetic diversity than what is found outside of Africa (48–50). This diversity largely mirrors the linguistic substructure and complexity that is found across African populations and is informed by mass human migration and admixture events. For example, Southern African populations cluster distinctly from other Bantu speaking populations in the more northern regions of Africa due to admixture with pre-existing populations like the San and Khoi. Thus, it remains to be seen what population specific architecture exists in the IG loci across different African populations

Immunogenetic sequencing techniques

There are a variety of techniques to examine genetic diversity across the IG loci. Amplicon based NGS (Next Generation Sequencing) approaches have been utilised to examine diversity in IGHG and IGHV genes from non-European populations to reveal significant genetic diversity (38,46,47). Larger scale sequencing methods have been developed to sequence IG loci. Initially these were fosmid and bac clone methods, with long-read sequencing of IG loci conjugated into these constructs (35,45). However, modern capture-based methods have also been developed which use magnetic beads that bind specifically to large fragments of DNA from IG loci (39,42,43). These long-read sequencing methods allow for the analysis of both polymorphic and structural diversity, uncovering variation that exists outside of IG genes. This is further discussed in Chapter 1.

Outside of genomic techniques, the genetic diversity of IG genes can also be studied by sequencing antibody transcripts. In a naive antibody repertoire, from IgM B-cell repertoires, the germline IG genes that produce the repertoire can be inferred (37). This approach also allows for the examination of expression biases within the repertoire, where certain IG genes are preferentially expressed over other genes (37). However, this approach has some important limitations. The inference of potentially novel alleles can be false due to the inherent mutations introduced during somatic hypermutation, although this can be partially abrogated by using low mutated IgM sequences. Transcriptomic based characterisation of IG genes also hinges on the expression of those genes, if a gene is deleted or expressed at very low levels this can affect the analysis for that gene. Furthermore, most sequencing of antibody transcripts is performed on short-read sequencing platforms, such as the Illumina MiSeq, which can only cover the VDJ region and a small portion of the constant gene (36,51,52). Thus, genetic diversity within constant genes is not well represented in most antibody transcript datasets. There are long-read methods for sequencing expressed antibodies, but they are not often utilized in favor of the depth from short-read sequencing platforms (53). High sequencing depth is necessary to be able to fully categorise an antibody response.

IG Gene Nomenclature

Here we describe the norms of describing antibody genes that are standard within the immunogenetic community. A description of an allele may look like the following: IGHV1-2*01. IGH denotes the locus of origin for the gene. IGHV describes the class of the gene. The phylogenetic subgroup, for V, D and J genes only, is delineated by IGHV1. Thus, the full gene name is IGHV1-2, where 01 is an allele of IGHV1-2. This naming convention is different for

the constant genes, where a gene is described in this manner: IGHA1. The isotype that is encoded by this gene is IgA1 and an allele of this gene would be described as IGHA1*01. It should also be noted that gene and allele names are not italicised within the field of immunogenetics, in contrast to what is the standard when discussing other genes. An allele can refer to a single SNP or a haplotype of SNPs which differentiate various alleles of a gene.

	VDJ Genes	Constant Genes
Genetic Locus:	IGH	IGH
Gene Class:	IGHV	IGHA
Phylogenetic Subgroup:	IGHV1	
Gene Name:	IGHV1-2	IGHA1
Full Allele Name:	IGHV1-2*01	IGHA1*01

Figure 4: Breakdown of IG gene naming conventions.

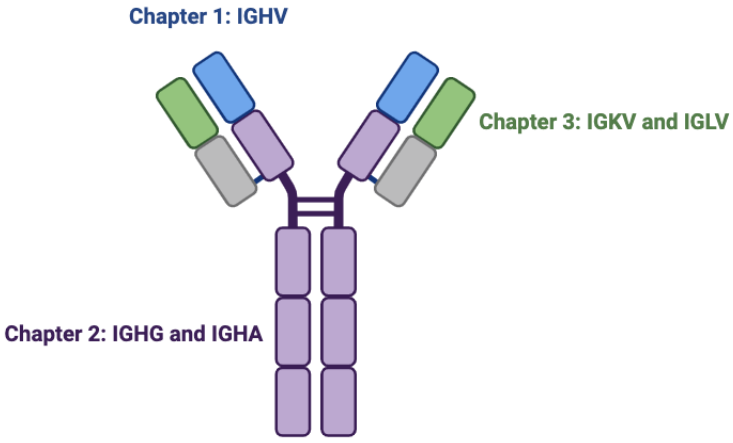
The Antibody Response in HIV

While this thesis will focus predominantly on the germline genetic diversity in the IG loci, there will be contextual discussion concerning infectious diseases, especially HIV. The antibody response in HIV is specific due to the chronic nature of the infection and the high mutability of the virus (13). This creates a dynamic in people living with HIV where the immune system and the host's viral population evolve in tandem as the virus escapes antibody responses through mutation and the antibody responses adjust to counter this escape (54,55). Over time, this coevolution can produce antibodies with unique traits; such as germline gene usage, amino acid motifs or CDR lengths; that can neutralise multiple different strains of HIV. These are termed broadly neutralising antibodies (bNAbs). Broadly neutralising antibody lineages often have very particular features that are shared across lineages that target the same epitope. For example, bNAbs that target the CD4 binding site on the HIV viral envelope (such as VRC01) are often derived from the germline IGHV1-2 gene and exhibit a high degree of affinity maturation with a double amino acid deletion in the light chain (56).

Germline targeting vaccine strategies are in development as potential preventative measures against HIV (57,58). These aim to elicit potent and broad responses that can protect against the wide array of circulating HIV strains by eliciting the creation of specific bNAb lineages. The germline immunogen is designed with high affinity for potential precursors of a bNAb lineage, which share the same germline genes and often specific germline-encoded residues in the antigen-binding region of the antibody. An example of this approach is seen with the prototype antigen nanoparticle eOD-GT8 (59). In human testing, it was found that 26 of the 27 donors produced VRC01-like precursor antibodies (60). The single participant that failed to produce any VRC01-like precursors entirely lacked either IGHV1-2*02 or IGHV1-2*04 in contrast to the other participants. This example demonstrates that in targeting specific germline IG genes, the genetic composition of the germline plays a critical role. Furthermore, it is likely that germline variation that affects the relative expression of IG genes will also have an impact on germline targeting vaccine responses. Thus, it is vital to characterise IG germline gene variation across different populations and elucidate the relationship between genetic variation and the expressed antibody repertoire to better inform these germline targeting vaccine approaches.

Chapter Overview

This thesis examines several different types of IG genes within a South African population, characterising novel genetic diversity and examining the potential impact of genetic variation in IG genes on the functional antibody response.



Chapter 1: Novel polymorphic and copy number diversity in the antibody IGH locus of South African Individuals

This chapter contains a published journal article (in Immunogenetics) that examines heavy chain variable (IGHV) genetic diversity within a South African population utilising both transcriptomic and genomic approaches.

Chapter 2: Identification of novel IgA and IgG antibody alleles in a South African population

This chapter explores genetic diversity in IGHA1 which encodes most IgA antibodies, as well as the genes that encode IgG subclasses - IGHG1, IGHG2, IGHG3 and IGHG4. This involved utilising a targeted amplicon based NGS approach to study these genes.

Chapter 3: Meta-analysis of IGLV and IGKV repertoires from South African and European populations

This chapter examines light chain expressed repertoires from three different populations, inclusive of this South African population. The aim of this was to assess dynamics of gene usage bias within IGKV and IGLV repertoire between different populations and what link there may be between the genetics of the germline and gene usage in the expressed repertoire.

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The Generation of Antibody Diversity [Internet]. Nih.gov. Garland Science; 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26860/

2. Janeway CA, Travers P, Walport M, Shlomchik MJ. Immunobiology [Internet]. Nih.gov. Garland Science; 2001. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10757/

3. Boyd SD, Jackson KJL. Predicting vaccine responsiveness. Cell Host Microbe. 2015;17(3):301–7.

4. Pollard AJ, Bijker EM. A guide to vaccinology: from basic principles to new developments. Nat Rev Immunol. 2021 Feb;21(2):83–100.

5. Lefranc MP, Lefranc G. The Immunoglobulin Factsbook. 2001. 72–239 p.

6. Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies. 2019 Dec 3;8(4):55.

7. Pomarici ND, Cacciato R, Kokot J, Fernández-Quintero ML, Liedl KR. Evolution of the Immunoglobulin Isotypes—Variations of Biophysical Properties among Animal Classes. Biomolecules. 2023 May 8;13(5):801.

8. Liu J, Wang Y, Xiong E, Hong R, Lu Q, Ohno H, et al. Role of the IgM Fc Receptor in Immunity and Tolerance. Front Immunol. 2019 Mar;10:529.

9. Ben Mkaddem S, Rossato E, Heming N, Monteiro RC. Anti-inflammatory role of the IgA Fc receptor (CD89): from autoimmunity to therapeutic perspectives. Autoimmun Rev. 2013 Apr;12(6):666–9.

10. Leusen JHW, Nimmerjahn F. The Role of IgG in Immune Responses. In: Nimmerjahn F, editor. Molecular and Cellular Mechanisms of Antibody Activity. New York, NY: Springer New York; 2013. p. 85–112.

11. Shamji MH, Valenta R, Jardetzky T, Verhasselt V, Durham SR, Würtzen PA, et al. The role of allergen-specific IgE, IgG and IgA in allergic disease. Allergy. 2021;76(12):3627–41.

12. Gutzeit C, Chen K, Cerutti A. The enigmatic function of IgD: some answers at last. Eur J Immunol. 2018 Jul;48(7):1101–13.

13. Mascola JR, Haynes BF. HIV-1 neutralizing antibodies: understanding nature's pathways. Immunol Rev. 2013 Jul;254(1):225–44.

14. Lu LL, Suscovich TJ, Fortune SM, Alter G. Beyond binding: antibody effector functions in infectious diseases. Nat Rev Immunol. 2018 Jan;18(1):46–61.

15. Chung A, Rollman E, Johansson S, Kent SJ, Stratov I. The utility of ADCC responses in HIV infection. Curr HIV Res. 2008 Nov;6(6):515–9.

16. Tay MZ, Wiehe K, Pollara J. Antibody-Dependent Cellular Phagocytosis in Antiviral

Immune Responses. Front Immunol. 2019 Feb;10:332.

17. Damelang T, de Taeye SW, Rentenaar R, Roya-Kouchaki K, de Boer E, Derksen NIL, et al. The Influence of Human IgG Subclass and Allotype on Complement Activation. J Immunol Author Choice. 2023 Dec 1;211(11):1725–35.

18. Shields RL, Namenuk AK, Hong K, Meng YG, Rae J, Briggs J, et al. High resolution mapping of the binding site on human IgG1 for Fc gamma RI, Fc gamma RII, Fc gamma RIII, and FcRn and design of IgG1 variants with improved binding to the Fc gamma R. J Biol Chem. 2001 Mar 2;276(9):6591–604.

19. Talathi SP, Shaikh NN, Pandey SS, Saxena VA, Mamulwar MS, Thakar MR. FcγRIIIa receptor polymorphism influences NK cell mediated ADCC activity against HIV. BMC Infect Dis. 2019 Dec 16;19(1):1053.

20. de Taeye SW, Bentlage AEH, Mebius MM, Meesters JI, Lissenberg-Thunnissen S, Falck D, et al. FcγR Binding and ADCC Activity of Human IgG Allotypes. Front Immunol. 2020 May 6;11:740.

21. Richardson SI, Lambson BE, Crowley AR, Bashirova A, Scheepers C, Garrett N, et al. IgG3 enhances neutralization potency and Fc effector function of an HIV V2-specific broadly neutralizing antibody. PLoS Pathog. 2019 Dec;15(12):e1008064.

22. Chi X, Li Y, Qiu X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology. 2020 Jul;160(3):233–47.

23. Sadofsky MJ. The RAG proteins in V(D)J recombination: more than just a nuclease. Nucleic Acids Research. 2001 Apr 1;29(7):1399–409.

24. Peters JM. How DNA loop extrusion mediated by cohesin enables V(D)J recombination. Curr Opin Cell Biol. 2021 Jun 1;70:75–83.

25. Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. Microbiol Spectr. 2014 Dec;2(6):10.1128/microbiolspec.MDNA3-0041–2014.

26. Zhang Y, Zhang X, Ba Z, Liang Z, Dring EW, Hu H, et al. The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. Nature. 2019 Sep;573(7775):600–4.

27. Boyd SD, Joshi SA. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Microbiol Spectr. 2014 Oct 10;2(5):10.1128/microbiolspec.aid-0017–2014.

28. Feng Y, Seija N, Di Noia JM, Martin A. AID in Antibody Diversification: There and Back Again. Trends Immunol. 2020 Jul;41(7):586–600.

29. Tang C, Bagnara D, Chiorazzi N, Scharff MD, MacCarthy T. AID Overlapping and Polη Hotspots Are Key Features of Evolutionary Variation Within the Human Antibody Heavy Chain (IGHV) Genes. Front Immunol. 2020 Apr;11:788.

30. Stavnezer J, Guikema JEJ, Schrader CE. Mechanism and regulation of class switch recombination. Annu Rev Immunol. 2008;26:261–92.

31. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 2012;13(5):363–73.

32. Collins AM, Watson CT. Immunoglobulin Light Chain Gene Rearrangements, Receptor Editing and the Development of a Self-Tolerant Antibody Repertoire. Front Immunol. 2018 Oct;9:2249.

33. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. 2012;569–604.

34. Watson CT, Glanville J, Marasco WA. The Individual and Population Genetics of Antibody Immunity. Trends Immunol. 2017 Jul;38(7):459–70.

35. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. Am J Hum Genet. 2013 Apr;92(4):530–46.

36. Vázquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M, et al. High-Quality Library Preparation for NGS-Based Immunoglobulin Germline Gene Inference and Repertoire Expression Analysis. Front Immunol. 2019;10:660.

37. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, et al. Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. Front Immunol. 2019 Mar;10:435.

38. Scheepers C, Shrestha RK, Lambson BE, Jackson KJL, Wright IA, Naicker D, et al. Ability To Develop Broadly Neutralizing HIV-1 Antibodies Is Not Restricted by the Germline Ig Gene Repertoire. J Immunol. 2015;194(9).

39. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. Front Immunol. 2020 Sep;11:2136.

40. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nat Commun. 2019;10(1):628.

41. Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, et al. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. Nucleic Acids Res. 2020 Jun;48(10):5499–510.

42. Gibson WS, Rodriguez OL, Shields K, Silver CA, Dorgham A, Emery M, et al. Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. Genes Immun. 2023 Feb;24(1):21–31.

43. Engelbrecht E, Rodriguez OL, Shields K, Schultze S, Tieri D, Jana U, et al. Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. Genes Immun. 2024 Aug;25(4):297–306.

44. Mikocziova I, Peres A, Gidoni M, Greiff V, Yaari G, Sollid LM. Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes. iScience. 2021 Sep;103192.

45. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, et al. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals

locus-specific signatures of genetic diversity. Genes Immun. 2015 Jan;16(1):24–34.

46. Calonga-Solís V, Malheiros D, Beltrame MH, Vargas L de B, Dourado RM, Issler HC, et al. Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection. Front Immunol. 2019 Jun;10:1161.

47. Bashirova AA, Zheng W, Akdag M, Augusto DG, Vince N, Dong KL, et al. Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) genes. Genes Immun. 2021 Dec;22(7):327–34.

48. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020 Oct;586(7831):741–8.

49. Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K, et al. Genetic substructure and complex demographic history of South African Bantu speakers. Nature Communications. 2021 Apr 7;12(1).

50. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019 Jan;51(1):30–5.

51. Chaudhary N, Wesemann DR. Analyzing Immunoglobulin Repertoires. Front Immunol. 2018 Mar;9:462.

52. Prabakaran P, Glanville J, Ippolito GC. Editorial: Next-Generation Sequencing of Human Antibody Repertoires for Exploring B-cell Landscape, Antibody Discovery and Vaccine Development. Front Immunol. 2020;11:1344.

53. Ford M, Haghshenas E, Watson CT, Sahinalp SC. Genotyping and Copy Number Analysis of Immunoglobin Heavy Chain Variable Genes Using Long Reads. iScience. 2020 Mar;23(3):100883.

54. Doria-Rose NA, Landais E. Coevolution of HIV-1 and broadly neutralizing antibodies. Curr Opin HIV AIDS. 2019 Jul;14(4):286–93.

55. Bhiman JN, Anthony C, Doria-Rose NA, Karimanzira O, Schramm CA, Khoza T, et al. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. Nat Med. 2015 Nov;21(11):1332–6.

56. Zhou T, Georgiev I, Wu X, Yang ZY, Dai K, Finzi A, et al. Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01. Science. 2010 Aug 13;329(5993):811–7.

57. Haynes BF, Wiehe K, Borrow P, Saunders KO, Korber B, Wagh K, et al. Strategies for HIV-1 vaccines that induce broadly neutralizing antibodies. Nat Rev Immunol. 2023 Mar;23(3):142–58.

58. Moore PL, Williamson C. Approaches to the induction of HIV broadly neutralizing antibodies. Curr Opin HIV AIDS. 2016 Nov;11(6):569–75.

59. Cohen KW, De Rosa SC, Fulp WJ, deCamp AC, Fiore-Gartland A, Mahoney CR, et al. A first-in-human germline-targeting HIV nanoparticle vaccine induced broad and publicly targeted helper T cell responses. Sci Transl Med. 2023 May 24;15(697):eadf3309.

60. Leggat DJ, Cohen KW, Willis JR, Fulp WJ, deCamp AC, Kalyuzhniy O, et al. Vaccination induces HIV broadly neutralizing antibody precursors in humans. Science. 2022 Dec 2;378(6623):eadd6502.

61. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci Rep. 2016 Aug;6(1):20842.

Samples and Ethics

Ethics Declaration

Informed consent was obtained for the use of stored samples that were collected as part of the original CAPRISA 002 and 004 trials. CAPRISA and the use of their samples by AIRU/NICD have ethics clearance from the University of the Witwatersrand Human Research Ethics Committee (M210892) and from the University of Kwa-Zulu Natal Biomedical Research Ethics Committee (E013/04). Ethics clearance for this thesis was obtained from the University of the Witwatersrand Human Research Ethics Committee (M220242). The relevant certificates have been provided as Addendum A.

CAPRISA Participant Information

The various data in this thesis was produced by the sequencing the same set of participants, We outline here which participant samples were utilised in each section in the following table:

CAPRISA ID	**HIV Status**	**Sequenced in Scheepers *et. al (2015)***	**Chapter 1: IGHV**		**Chapter 2:**		**Chapter 3: IGK and IGLV**
			Genomic	Transcriptomic	IGHA	IGHG	
008	+	X	X	X	X		
040	+		X				
045	+		X		X		
062	-		X		X		
064	-		X				
065	+		X		X		
080	-		X				
085	+		X		X	X	
088	+	X	X	X	X	X	X
103	-		X				

105	-		X		X		
108	-		X		X		
109	-		X				
110	-		X				
115	-		X				
117	-		X		X		
118	-		X		X		
119	-		X		X		
120	-		X				
121	-		X				
137	+	X	X		X		
145	-		X		X		
146	-		X				
147	-		X		X		
148	-		X		X		
161	-		X		X		
165	-		X		X		
172	-		X		X		
177	+	X	X		X	X	
206	+	X		X	X	X	X
220	+		X		X		
221	+	X	X				
222	+		X		X		
225	+		X		X		
228	+		X	X	X	X	
229	+	X	X	X	X		
244	+	X	X	X	X		
248	+	X	X	X	X		X
250	-		X				
252	-		X				
253	-		X		X		

254	-		X		X		
255	+		X	X	X	X	X
256	+	X		X		X	X
257	+	X	X	X	X	X	X
268	+	X	X				
280	+		X		X		
286	+		X				
287	+	X	X		X		
288	+		X		X		
290	+	X	X				
292	+	X	X				
295	+	X	X		X		
301	+		X		X		
306	+		X		X		
310	+	X	X		X		
312	+	X	X	X	X		
314	+	X	X	X	X		
326	+		X		X		
331	+		X		X		
335	+		X		X		
341	+		X		X		
342	+		X		X		
346	+		X		X		
349	+		X				
357	+		X				
363	+		X				
365	+		X				
378	+		X		X		
379	+		X				
383	+		X				
387	+		X				

Chapter 1:

Novel polymorphic and copy number diversity in the antibody IGH locus of South African Individuals

---

Originally published online in *Immunogenetics* on 04 December 2024, full citation for this article is:

Marsden AA, Corcoran M, Karlsson Hedestam G, Garrett N, Karim SSA, Moore PL, et al. Novel polymorphic and copy number diversity in the antibody IGH locus of South African individuals. Immunogenetics. 2024 Apr;77(1):6.

Licensing and Permissions to reproduce this article as part of this thesis:

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

The supplementary figures and tables can be found under Addendum B on pages

Immunogenetics (2025) 77:6
https://doi.org/10.1007/s00251-024-01363-7

ORIGINAL ARTICLE

Novel polymorphic and copy number diversity in the antibody IGH locus of South African individuals

Alaine A. Marsden[1,2] · Martin Corcoran[3] · Gunilla Karlsson Hedestam[3] · Nigel Garrett[4,5] · Salim S. Abdool Karim[4,6] · Penny L. Moore[1,2,4] · Dale Kitchin[1,2] · Lynn Morris[1,4] · Cathrine Scheepers[1]

Received: 31 May 2024 / Accepted: 19 November 2024
© The Author(s) 2024

Abstract
The heavy chain of an antibody is crucial for mediating antigen binding. IGHV genes, which partially encode the heavy chain of antibodies, exhibit vast genetic diversity largely through polymorphism and copy number variation (CNV). These genetic variations impact population-level expression levels. In this study, we analyzed expressed antibody transcriptomes and matched germline IGHV genes from donors from KwaZulu-Natal, South Africa. Amplicon NGS targeting germline IGHV sequences was performed on genomic DNA from 70 participants, eight of whom had matched datasets of expressed antibody transcriptomes. Germline IGHV sequencing identified 161 unique IGHV alleles, of which 32 were novel. A further 21 novel IGHV alleles were detected in the expressed transcriptomes of these donors. We also examined the datasets for CNV, uncovering gene duplications of 10 IGHV genes from germline sequencing and 33 genes in the expressed transcriptomes. Many of the IGHV gene duplications have not been described in other populations. This study expands our understanding of genetic differences in distinct populations and suggests the potential impact of genetic diversity on immune responses.

Keywords Immunogenetics · Antibodies · Human genetics · Immunology

Introduction

Antibodies, or immunoglobulins, are a fundamental component of the adaptive immune response. The genes that encode the heavy chain of an antibody are all found within the immunoglobulin heavy chain (IGH) locus on chromosome 14 (Lefranc and Lefranc 2001). These include the variable (IGHV), diversity (IGHD), and joining (IGHJ) genes that are joined together through VDJ recombination to create the antigen-binding site. The IGH locus is one of the most polymorphic and structurally diverse loci in the human genome (Watson and Breden 2012). Single nucleotide variants (SNVs) across the locus create allelic diversity for IGH genes and large structural variants produce copy number variation (CNV) through the insertion, deletion, and duplication of multiple genes at a time (Watson et al. 2013). Amongst IGH genes, IGHV genes are the most numerous in the IGH locus, comprising up to 50 functional genes, 6 open reading frames, and 81 pseudogenes, divided into eight phylogenetic subgroups (IGHV1-8) (Lefranc and Lefranc 2001). For the purposes of this study, we focus on functional IGHV genes (IGHV1-7), avoiding pseudogenes and open reading frames.

✉ Cathrine Scheepers
cathrine.scheepers@wits.ac.za

1 SA MRC Antibody Immunity Research Unit (AIRU), University of the Witwatersrand, Johannesburg, South Africa

2 Centre for HIV and STIs, HIV Virology Section, National Institute for Communicable Diseases (NICD), a Division of the National Health Laboratory Service (NHLS), Johannesburg, South Africa

3 Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

4 Centre for the AIDS Programme of Research in South Africa (CAPRISA), University of KwaZulu-Natal, Durban, South Africa

5 Discipline of Public Health Medicine, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa

6 Department of Epidemiology, Mailman School of Public Health, Columbia University, Columbia, NY, USA

Published online: 04 December 2024

Previous studies have demonstrated that polymorphic diversity and structural variation exhibit population-specific patterns (Watson et al. 2013; Avnir et al. 2016; Rodriguez et al. 2020). For example, the duplication of IGHV1-69 is more prevalent in African populations than in East Asian populations (Avnir et al. 2016). This is important because there is evidence that germline IGHV variation affects the expressed antibody transcriptomes, with CNV or polymorphism impacting IGHV gene usage (Avnir et al. 2016; Kenter et al. 2021; Pennell et al. 2023; Mikocziova et al. 2021a). Alterations in the expressed transcriptome frequently occur when an IGHV gene is involved in a duplication or complex event that alters copy number (such as for IGHV1-69, IGHV3-64D, IGHV5-10–1, IGHV1-8, and IGHV3-9, IGHV3-23) (Avnir et al. 2016; Rodriguez et al. 2023). Furthermore, SNVs in intergenic regions have been associated with alterations in gene expression across several genes, although the mechanism for this is unclear. For example, an A/G SNV occurring within the IGH locus (rs8008062), occurring 120 Kbp from any IGHV gene or known structural variant, has been shown to affect the expression of seven germline IGHV genes (Rodriguez et al. 2023). Of the seven genes, four (IGHV4-31, IGHV3-53, IGHV4-61, and IGHV1-69) were expressed at higher levels with a G nucleotide at the position as opposed to an A. Conversely, the remaining three genes (IGHV4-61, IGHV3-64, and IGHV3-66) were expressed at lower levels when a G nucleotide was found at the position in place of an A (Rodriguez et al. 2023). This study suggests that the SNV occurs in a regulatory region that modulates the chromatin formation of the locus, affecting the accessibility of IGHV genes to various enzyme complexes.

IGHV gene usage biases have also been described in a variety of infection and vaccination responses, resulting in convergent antibody responses across individuals. Influenza vaccine and infection responses frequently use the IGHV3-7, IGHV1-69, and IGHV4-39 genes. In SARS-CoV-2, there is evidence that IGHV3 family genes are commonly used in antibody responses (Nielsen et al. 2020; Mor et al. 2021). For HIV-1, broadly neutralizing antibody lineages often display the same IGHV gene usage across different epitopes, such as IGHV1-69 (MPER), IGHV4-34 (V3-Glycan), IGHV1-2 (CD4bs and V3-Glycan), and IGHV3-15 (MPER and V2-glycan) (Moyo et al. 2020; Doria-Rose et al. 2014; Richardson and Moore 2019; Zhou et al. 2015; Soto et al. 2016; Walker et al. 2009). If an IGHV gene is frequently observed in the transcriptomes of multiple individuals that have mounted a response against a pathogen or immunogen, it follows that the gene likely produces precursor B-cells with high affinity. Thus, having a gene duplication or genetic mutation that leads to overexpression of that particular IGHV gene may be beneficial in mounting a response.

In addition to gene usage biases, previous studies have shown that particular IGHV alleles can lead to more potent and/or broad responses to pathogens (Mikocziova et al. 2021a; Safonova et al. 2022). This has been demonstrated in influenza where there is a strong bias in the usage of IGHV1-69 and also key IGHV1-69 alleles that have a phenylalanine at position 54 produce more potent neutralizing responses (Avnir et al. 2016; Wheatley et al. 1950; Pappas et al. 2014; Jackson et al. 2014; Avnir et al. 2014). In HIV, a recent study has shown that VRC01-like CD4bs antibodies are produced more frequently in participants that express IGHV1-2*02 (Vaccine genetics of IGHV1-2 VRC01-class broadly neutralizing antibody precursor naïve human B cells - PubMed [Internet]. 2024).

We have previously sequenced the V-regions of functional IGHV genes in 28 participants from the CAPRISA cohorts (Scheepers et al. 2015). Here, we expand the scope of sequencing IGHV genes to include the regulatory parts of the gene, including the leader sequence and the recombination signal sequence (RSS). We developed an amplicon-based NGS approach to sequence the full IGHV gene from germline DNA, to identify novel IGHV alleles and genotypes for CNV. Furthermore, we mined existing transcriptomic data, sequenced from naive IgM B-cells of matched donors, to explore genetic diversity and its link to gene expression. Haplotype reconstruction was utilized from the transcriptomic data to examine CNV. Overall, we used a combination of both germline sequencing and transcriptome sequencing, within matched donors, to evaluate the impact of germline variation on the transcriptome.

Methods and materials

Study participants

This study utilized samples from the CAPRISA 002 and 004 cohorts. These cohorts consisted of black African women recruited in the KwaZulu-Natal province of South Africa. CAPRISA 002 examined acute HIV infection (Loggerenberg et al. 2008), while CAPRISA 004 was a clinical trial of tenofovir gel in preventing HIV (Abdool Karim et al. 2010). There is overlap between the participants that were involved in both studies, thus representing a combined cohort. A total of 70 women from both CAPRISA cohorts participated in this study, of whom 58 were participants living with HIV, including the 28 donors in the previous study by Scheepers et al. (2015), and 12 were participants without HIV. Informed consent was obtained from all participants for sample collection and storage for future studies.

Amplification of germline IGHV genes

Genomic DNA was extracted from stored peripheral blood mononuclear cells (PBMC) using the Wizard® Genomic DNA Purification Kit (Promega; Madison, WI, USA). Primers were designed to amplify the full IGHV gene, including the regulatory regions, for the first six IGHV subgroups (Supplementary Table 1). The IGHV7 subgroup of genes failed to yield full-length amplicons and was omitted from this study. However, we have previously reported on IGHV7 in this population, which was done utilizing a different set of primers designed to target within the V-region of the IGHV gene (Scheepers et al. 2015). The primers used in this study differ in that they bind outside the entire IGHV gene, in the untranslated regions because our aim was to produce full-length sequences. The forward primer for each subgroup bound before leader part 1 of the respective IGHV gene and the reverse primer bound after the RSS. Amplicon sizes ranged from ~ 570 bp for IGHV1 genes to ~ 880 bp for IGHV4 genes. Each PCR reaction contained 15.75 µL of molecular grade water, 5 µL of 5 × Platinum SuperFi Buffer, 0.5 µL of a dNTP mix (10 µM each), 1.25 µL of forward primer (10 µM), 1.25 µL of reverse primer (10 µM), and 0.25 µL of Platinum SuperFi Polymerase (Thermo Fisher Scientific; Waltham, MA, USA) (2.5 U/µL). To this mix, 1 µL of DNA (10 ng/µL) was added. The cycling conditions had an initial denaturation of 30 s at 98 °C. There were 35 cycles of denaturation at 98 °C for 30 s, annealing for 10 s, and extension at 72 °C for 1 min. The final extension was for 2 min at 72 °C. Annealing temperatures varied by primer pair (see Supplementary Table 1). Each sample was run in duplicate to ensure adequate coverage and minimize primer biases.

NGS library preparation and sequencing of germline IGHV genes

The IGHV amplicons were purified using AMPure XP beads (Beckman Coulter; Brea, CA, USA). Beads were added to the amplicons at a 1:0.75 ratio, to ensure the removal of small DNA fragments (< 200 bp). IGHV1 amplicons were sequenced on the Illumina MiSeq platform due to MiSeq being more appropriate for the amplicon size (570 bp) and the greater sequencing depth that can be produced in a run. This was necessary due to the large number ($n = 21$) of genes in this subgroup and the high allelic diversity of IGHV1 genes, such as IGHV1-69. The other subgroups were sequenced using the PacBio Sequel platform (IGHV2 through IGHV6); these were then pooled using a gene-appropriate ratio (Supplementary Table 1), where the final concentration for each subgroup was proportional to the number of genes in each reaction. All amplicons were indexed with participant-specific molecular barcodes. IGHV1 amplicons, sequenced on the Illumina MiSeq, were indexed with the Nextera Flex indices (Illumina; San Diego, CA, USA) using the Failsafe PCR kit (Epicentre; Madison, WI, USA) in a limited 11 cycle PCR. Each indexing reaction contained 10 µL of Nextera index mix, 25 µL of Failsafe PCR 2× PreMix, 13.5 µL of water, 0.5 µL Failsafe PCR enzyme mix, and 1 µL of amplicon DNA. IGHV2 to IGHV6 were indexed using Pacific Biosciences (Menlo, CA, USA) barcodes and the Platinum SuperFi kit, using the same protocol as initial amplification except with the indexes as primers.

Sequencing library quality assessment and sequencing were performed by the NICD Sequencing Core. Briefly, the amplicon library concentration was measured on a Qubit® 3.0 Fluorometer (Invitrogen; Waltham, MA, USA) and the size of the amplicons were then visualized using the 4200 TapeStation (Agilent Technologies; Santa Clara, CA, USA). The purified amplicon libraries were normalized to 2 nM and then sequenced on the Illumina MiSeq platform using the Illumina MiSeq v3 kit with a 10% PhiX spike in, to obtain 2 × 300 bp paired-end sequences. The library preparation for PacBio libraries was done using the SMRTbell Template Prep Kit (v.1.0) (Pacific Biosciences). Sequencing was carried out on PacBio® Sequel system using SMRT cell 1 M with a movie time of 600 min, pre-extension time of 90 min using Sequel Binding Kit (v.2.1), Sequel Sequencing Plate (v.2.1) and loading was done by diffusion. The resulting subreads were demultiplexed and the circular consensus sequences (CCSs) were generated using the "ccs" command (0.9999 minimum predicted accuracy and 3 minimum passes) on SMRT Link (v.6.0) (Pacific Biosciences).

Sequence analysis of germline IGHV sequences

Paired-end MiSeq FASTQ files were paired using PEAR (v0.9.8) (Zhang et al. 2014). PacBio CCS FASTQ files did not require this step. Primers were then removed from all reads, from both platforms, using CutAdapt (v3.5) (Martin 2011). The FASTQ files were converted to FASTA using the FASTX toolkit (Lab 2010). Removal of low-quality reads (Q score < 33), dereplication, and removal of low-abundance reads were done using VSEARCH (v2.22.1) (Rognes et al. 2016). Processed sequences were assigned gene and allele annotations using a custom Julia script developed for this analysis utilizing modules and packages from BioJulia (https://github.com/NICD-AIRU/AMarsden_IGHV_Analysis_2024). Briefly, the V-region of the IGHV genes were compared to a custom database, created using reference sequence data from IMGT (imgt.org, accessed in Jan 2023) (Lefranc et al. 2015; Manso et al. 2022; Lefranc 2014) and IgPdb (https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb), using BLAST (v2.12). The BLAST output was then parsed to select the closest match with the highest sequence identity for the sequence and assign a gene and allele for

the sequence. Pairwise alignment was then performed to identify SNVs (single nucleotide variants) in the regulatory regions of the gene using reference leader, intron, and RSS sequences from IMGT, noting the region and frequency of SNVs across the gene. The script also examined the number of unique alleles per person for each gene; if there were more than two unique alleles assigned to a gene, that gene was assumed to be duplicated in that participant.

Analysis of antibody transcriptome datasets

Of the 70 CAPRISA donors used for the germline IGHV sequencing, eight had donor-matched bulk IgM transcriptomic sequencing, obtained pre-infection, that were utilized in this study. Amongst the eight, two have been published in a study examining public clonotypes in longitudinal samples from HIV-infected individuals (Setliff et al. 2018). The sequence data from the previously published study was produced with the methodology outlined in the original paper (Setliff et al. 2018). Briefly, unbiased amplification of VDJs was performed using a multiplex approach, with forward primers designed to amplify all heavy chain regions by targeting the framework-1 region. These underwent quality assessment before sequencing on an Illumina MiSeq. The unpublished data in this study was produced using a 5′ multiplex method to amplify full VDJs, using a different primer set which targets the leader sequence (Vázquez Bernat et al. 2019).

All transcriptomic sequencing data were processed from paired-end Illumina MiSeq reads that were paired using PEAR. Sequence data was identified and annotated using PyIR, a python wrapper for IgBlast (v1.20.0) (Soto et al. 2020). Novel allele inference was performed using IgDiscover (v1.01) (Corcoran et al. 2016). Individual haplotypes were reconstructed with RabHit, based on the allele assignment and inference of novel alleles, initially made by IgDiscover using IGHJ6 as a gene anchor, which was heterozygous in 4/8 participants and yielded reliable results (results with confidence scores above 3 according to RabHit (v0.2.5) (Peres et al. 2019). Other IGHJ genes, IGHJ1-IGHJ5, were assessed, but fewer participants were heterozygous with sufficient expression for RabHit for these genes (IGHJ1:0, IGHJ2:0, IGHJ3:0, IGHJ4:2, IGHJ5:0). Custom Julia scripts (https://github.com/NICD-AIRU/AMarsden_IGHV_Analysis_2024) were used to compare the outputs of the transcriptomic and germline sequencing analyses. In brief, the scripts count the number of unique alleles per haplotype. The expressed allele calls were compared with the germline allele calls for the participants to establish concordance. The novel IgDiscover calls are compared to existing novel alleles from the germline sequencing and IgPDb alleles. We assessed the impact of duplication on expression by comparing unique VDJ counts between haplotypes with duplication to haplotypes without duplication, on a per gene basis. This was conducted on a gene-wise basis to avoid the effect of variable expression between genes. Duplications were also excluded from this analysis if they had a read count under 50 or if the Bayesian significance factor (k-value) was under 3.6, deeming the inference as unsubstantial (see Supplementary Table 7).

Results

Novel alleles identified by germline and bulk IgM transcriptome sequencing

Full-length germline sequencing of IGHV genes from 70 individuals revealed 161 alleles (Fig. 1A, Supplementary Table 2) across 33 IGHV genes. While 56% ($n=87$) of these alleles matched known sequences in the IMGT database (indicated in black), 21% ($n=32$) were novel (indicated in red). Novel sequences and their identity to known IMGT alleles are detailed in Supplementary Table 3. The majority of the remaining alleles (13%, $n=21$) matched those previously identified from the same donors, indicated in grey (Scheepers et al. 2015). The final 10% ($n=16$) of alleles matched inferred sequences from the IgPDb or OGRDB databases (indicated in blue). Unique alleles were determined using full-length sequences, inclusive of all regions of the IGHV gene. This created instances where V-regions between alleles were identical but there were differences in other regions such as the intron or leader sequence (Supplementary Table 4).

Some genes had less sequences than other genes, as evidenced by lower or non-existent allele calls, particularly those genes within the IGHV3 family. Due to the fact that we were able to amplify them in the transcriptomic data from the same donors (Fig. 1B) and have amplified them in previous germline sequencing (using different primers) (Scheepers et al. 2015), it is unlikely that the missing sequences result from germline deletions. Therefore, the missing sequences likely result from PCR biases due to primer multiplexing, with the primers favouring some genes over others. It is also possible that unknown nucleotide variation lays in the primer binding sites.

We next aimed to examine the expressed antibody transcriptome for further novel variants and determine if the novel alleles described in this study can be observed in expressed antibodies. From the larger cohort of 70 individuals, eight had matching bulk IgM-expressed datasets comprising more than 5 million total sequences. We performed germline inference to identify IGHV alleles in these expressed datasets (Fig. 1B, Supplementary Table 2). There were a greater number of IGHV genes observed using this approach, with alleles being identified

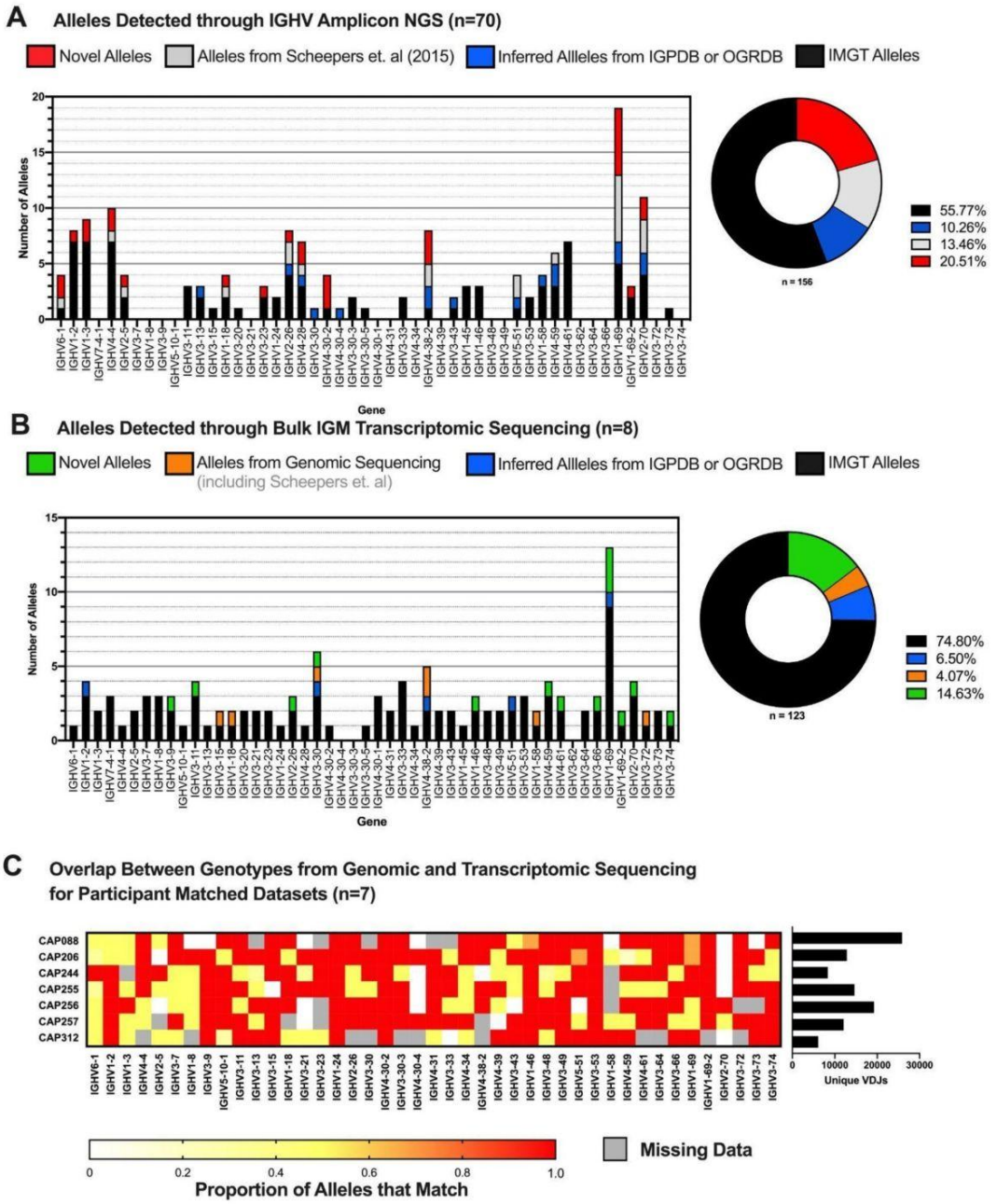
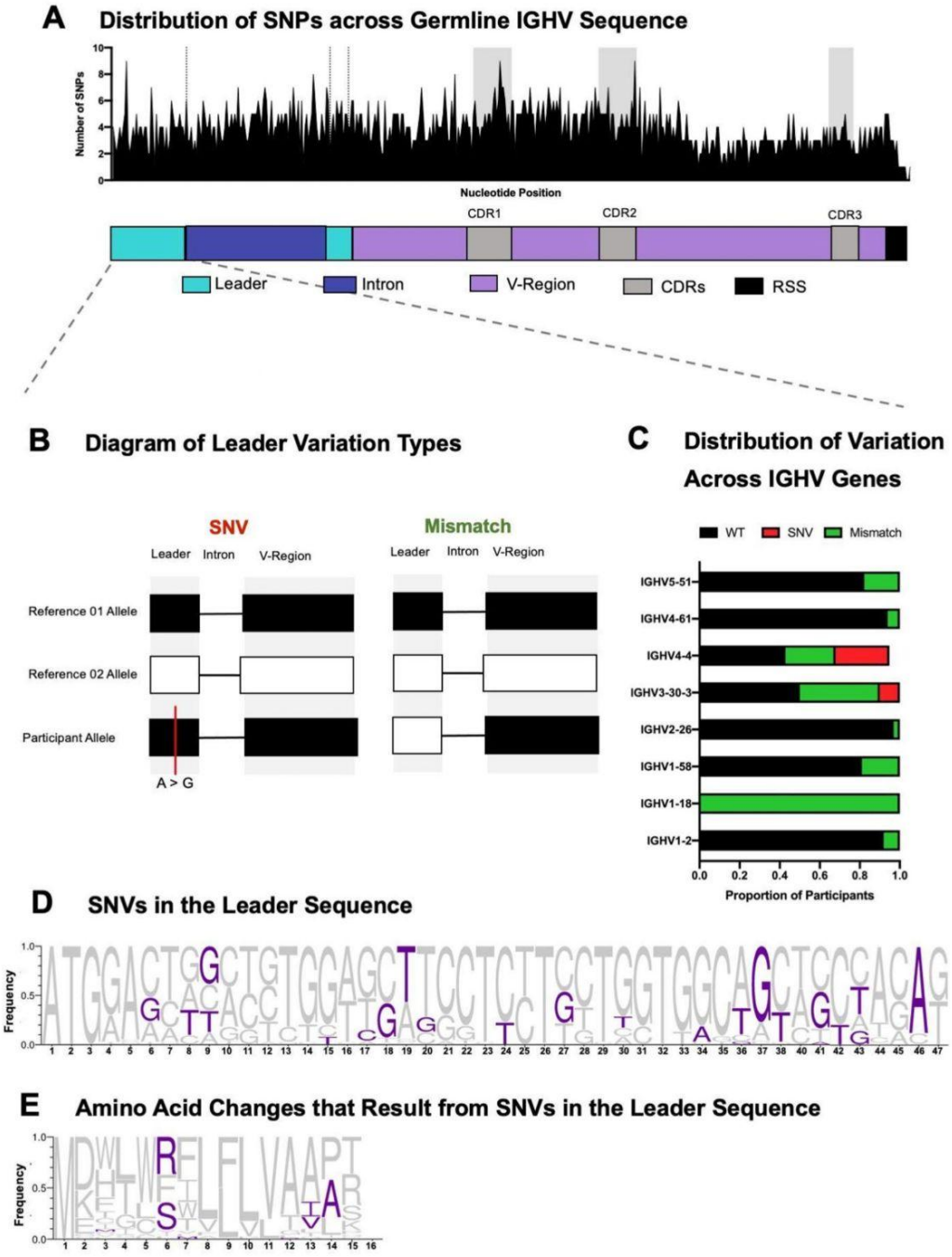


Fig. 1 Distribution of IGHV alleles identified through transcriptomic and germline sequencing. **A** Alleles identified from germline NGS, coloured by the alleles that were novel (red), previously identified in Scheepers et al. (2015) (grey), matched alleles in the IgPDb/OGRDB (blue), and IMGT (black) databases. **B** Alleles identified from transcriptome sequencing of bulk IgM, coloured by the categorization of the alleles with novel alleles (green), the alleles which match previous genomic sequencing from Scheepers et al. (2015) (orange), inferred alleles from other databases (blue), and IMGT alleles (black). **C** Heatmap showing the overlap between the genomic and transcriptomic sequencing datasets for seven matched participants. Participant 228 was excluded due to low VDJ count. Grey represents where either approach is lacking data for that gene. The sequencing depth from the transcriptome sequencing is represented by the bar graph to the right that shows the number of unique VDJs identified per participant



for 48/50 IGHV genes (compared to 33 in the germline data). In total, 123 alleles were identified; the majority (75%, $n = 92$) of which matched known alleles in the IMGT database (black, Fig. 1B). Most non-IMGT alleles were novel (green), making up 15% ($n = 18$) of alleles; identity to known IMGT alleles is shown in Supplementary Table 3. A small proportion (7%, $n = 8$, blue) of the sequences matched inferred alleles from other studies found in IgPDb or OGRDB (Lees et al. 2020), while a minority matched alleles from genomic sequencing in

◂**Fig. 2** Genetic diversity in the leader sequence of IGHV genes from germline sequencing. **A** SNPs across the full-length IGHV gene. The peaks represent the number of single nucleotide variations across the IGHV gene. The various regions of the gene are highlighted both on the graph and below in a schematic. It should be noted that CDR3 in this figure does not represent the whole CDR3 in a final antibody sequence as this is the germline IGHV portion of the CDR3, distinct from the CDR3 in the protein of an antibody that arises from the junction of the V, D, and J genes during somatic recombination. **B** Schematic representation of differing forms of leader part 1 sequence diversity; alignment is demonstrated in Supplementary Fig. 1. **C** Distribution of leader sequence diversity across specific IGHV genes with SNV in red and Mismatch in green. **D** Logogram representing the relative frequency of novel leader sequence variations when contrasted against reference leader sequence data (from IMGT). The novel bases are represented in purple. **E** Logogram of translated leader sequence data, representing novel amino acid changes (purple) as the result of leader sequence diversity

either this or the previously study on these donors (4%, $n = 5$, orange).

We compared the allele assignments for the donors that had high VDJ counts and found strong concordance between the germline and transcriptomic sets (Fig. 1C). This analysis focussed on genes where there was sufficient overlap between the datasets, excluding genes where there was substantial missing data. One of the participants (CAP228) was also excluded due to low VDJ counts. In genes where coverage from both techniques for a participant was high, concordance was strong (> 80%). The discrepancy between the datasets, where concordance was low, is partly due to large differences in IGHV coverage between the two methods. Furthermore, discordance may arise in calls between the datasets (the white boxes) as the result of the different way alleles are identified. In the transcriptome, not all IGHV alleles, or even certain IGHV genes, may be expressed, hampering the ability to efficiently discriminate true alleles from somatically mutated sequences. For example, IGHV2-70 is highly discordant in our dataset. This likely results from the low levels at which IGHV2-70 is expressed which makes it difficult to correctly infer alleles for this gene. For CAP312, there are two genomic novel alleles (IGHV6-1*03_C288T and IGHV6-1*01_C288T) which differ to the transcriptome (which shows IGHV6-1*01). It is possible that the genomic alleles contain mutations that prevent their expression or were not expressed at high levels at the particular time point sampled. Furthermore, the mutation lies towards the end of the transcript sequence where uncertainty from inference exists due to variable v sequence lengths. Thus, the discrepancy likely arises from technical limitations in the inference process. The primers utilized in the germline approach may sit in regions where unknown polymorphic diversity lays, affecting their efficacy, as opposed to the primers in the transcriptomic approaches which target conserved regions within the IGHV gene.

Leader sequence variation in full-length germline IGHV sequences

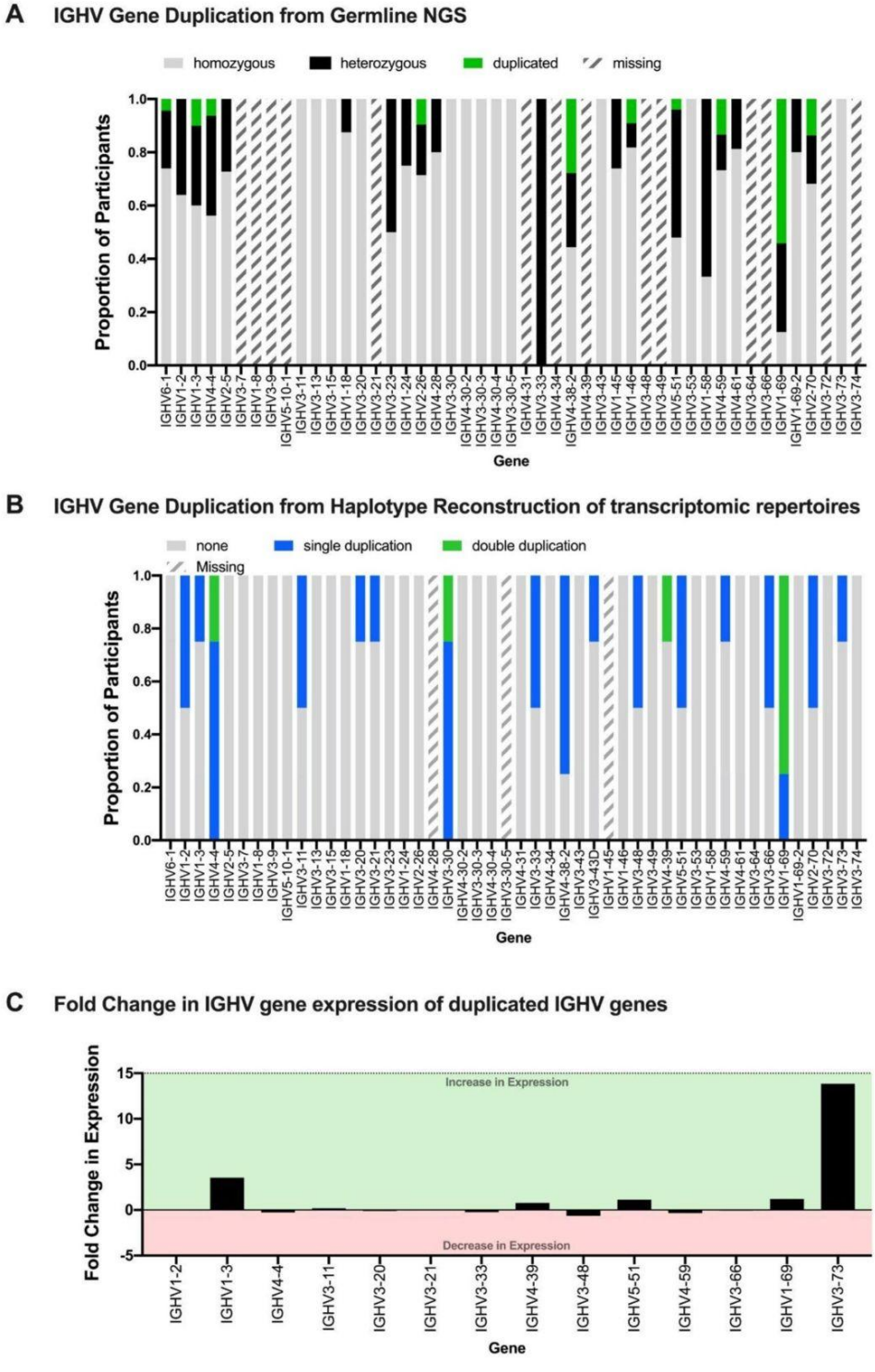
Using full-length germline sequences, we examined the distribution of nucleotide diversity across the whole gene, including the leader sequence and V-region (Fig. 2A). This revealed variation across all parts of the IGHV gene, not just in the V-region, which is the most extensively studied.

Within the first part of the leader region (before the intron, Supplementary Table 5), we identified two types of variation: one involving single nucleotide variation (SNVs) and the other in which the leader sequence is represented in IMGT but is associated with a different V-region than what has been detected in these sequences, resulting in a mismatch between the expected leader and V-regions. This latter variation resulted in a different assignment of the leader sequence from the V-region identity, which we termed mismatch (Fig. 2B). The mismatch may arise from nucleotide variation, such as SNVs, or these variants are the product of homologous recombination. The different variations are also represented in alignment form in Supplementary Fig. 1. These analyses excluded novel alleles and alleles that matched truncated data from IgPDb, as there is no reference leader sequence for these alleles for comparison to identify SNVs or mismatch. The mismatch variation was observed in 20% of alleles and present in eight genes, in contrast to the SNV which was detected in 6% of alleles and only present in two genes. Notably, all IGHV1-18 alleles had mismatched leaders.

Despite the variations being across different genes and alleles, there was commonality in the variable nucleotide positions (Fig. 2D). For several positions (nucleotides: A19T, A37G, and C46A) across the leader, novel mutations were more prevalent than the nucleotides from reference IMGT leader sequences. Several of these leader variants lead to amino acid changes (Fig. 2E), which may have implications for antibody expression that warrant further investigation.

Gene duplication confirmed by reconstructed haplotypes

CNV analysis was performed on both germline and transcriptomic datasets to examine the presence of gene duplication. For the germline NGS data (Fig. 3A), duplication was called when there were more than two unique alleles for a gene in a participant. From this, we estimated duplication in 10/33 IGHV genes. Gene deletion, conversely, was difficult to estimate with accuracy as a single unique allele may indicate homozygosity rather than a deletion. Furthermore, no alleles detected for a particular gene may be the result of poor sequencing coverage; thus, deletions were excluded from this analysis. A caveat of this approach is


A
IGHV Gene Duplication from Germline NGS
homozygous
heterozygous
duplicated
missing
Proportion of Participants
1.0
0.8
0.6
0.4
0.2
0.0
IGHV6-1
IGHV1-2
IGHV1-3
IGHV4-4
IGHV2-5
IGHV3-7
IGHV1-8
IGHV3-9
IGHV5-10-1
IGHV3-11
IGHV3-13
IGHV3-15
IGHV1-18
IGHV3-20
IGHV3-21
IGHV3-23
IGHV1-24
IGHV2-26
IGHV4-28
IGHV3-30
IGHV4-30-2
IGHV3-30-3
IGHV4-30-4
IGHV3-30-5
IGHV4-31
IGHV3-33
IGHV4-34
IGHV4-38-2
IGHV4-39
IGHV3-43
IGHV1-45
IGHV1-46
IGHV3-48
IGHV3-49
IGHV5-51
IGHV3-53
IGHV1-58
IGHV4-59
IGHV4-61
IGHV3-64
IGHV3-66
IGHV1-69
IGHV1-69-2
IGHV2-70
IGHV3-72
IGHV3-73
IGHV3-74
Gene
B
IGHV Gene Duplication from Haplotype Reconstruction of transcriptomic repertoires
none
single duplication
double duplication
Missing
Proportion of Participants
1.0
0.8
0.6
0.4
0.2
0.0
IGHV6-1
IGHV1-2
IGHV1-3
IGHV4-4
IGHV2-5
IGHV3-7
IGHV1-8
IGHV3-9
IGHV5-10-1
IGHV3-11
IGHV3-13
IGHV3-15
IGHV1-18
IGHV3-20
IGHV3-21
IGHV3-23
IGHV1-24
IGHV2-26
IGHV4-28
IGHV3-30
IGHV4-30-2
IGHV3-30-3
IGHV4-30-4
IGHV3-30-5
IGHV4-31
IGHV3-33
IGHV4-34
IGHV4-38-2
IGHV3-43
IGHV3-43D
IGHV1-45
IGHV1-46
IGHV3-48
IGHV3-49
IGHV4-39
IGHV5-51
IGHV3-53
IGHV1-58
IGHV4-59
IGHV4-61
IGHV3-64
IGHV3-66
IGHV1-69
IGHV1-69-2
IGHV2-70
IGHV3-72
IGHV3-73
IGHV3-74
Gene
C
Fold Change in IGHV gene expression of duplicated IGHV genes
Fold Change in Expression
15
10
5
0
-5
Increase in Expression
Decrease in Expression
IGHV1-2
IGHV1-3
IGHV4-4
IGHV3-11
IGHV3-20
IGHV3-21
IGHV3-33
IGHV4-39
IGHV3-48
IGHV5-51
IGHV4-59
IGHV3-66
IGHV1-69
IGHV3-73
Gene

◂**Fig. 3** Gene duplication in IGHV genes. **A** Duplication of IGHV genes detected in germline sequencing of 70 participants. The bars represent the proportion of the participants who were homozygous (grey), heterozygous (black), or were duplicated (green) for a given gene. **B** CNV detected from haplotype reconstruction of donor-matched expressed IgM data. The bars represent the proportion of the participants who contained no duplication (grey), a single haplotype duplication (blue), and duplications on both haplotypes that were resolved (green). **C** Fold change in expression between haplotypes with duplication and those without duplication. Expression was calculated by taking the difference between the relative proportions a particular gene was expressed as in association with either haplotype

that two unique alleles may exist on one chromosome while the other chromosome contains a deletion. Furthermore, this approach cannot take into account that duplications may be the result of completely identical alleles; thus, we could be underestimating duplications in this dataset. The participants that contained duplications and the alleles that contributed to duplication are noted in Supplementary Table 6. It should be noted that the analysis was performed on a per gene basis, examining participants that had coverage of that gene. Thus, the proportions do not represent a proportion of the total population but those that had sequences for a particular gene. For example, there were few participants that had IGHV3-33 sequences which resulted in two alleles identified (Fig. 1A); however, both of these participants were heterozygous for IGHV3-33.

Gene duplication has been reported in nine IGHV genes in other studies (Watson et al. 2013; Avnir et al. 2016). Of those genes, four (IGHV1-46, IGHV4-59, IGHV1-69, and IGHV2-70) were also duplicated in individuals in our population. The most commonly duplicated gene was IGHV1-69, which was duplicated in more than half of the population, while the other genes were less commonly duplicated (20%, $n=14$). Furthermore, six IGHV genes (IGHV6-1, IGHV1-3, IGHV4-4, IGHV2-26, IGHV4-38–2, and IGHV5-51) demonstrate novel duplication, not reported in other CNV studies. The most commonly duplicated gene amongst the novel duplications was IGHV4-38–2 (27%, $n=19$), showing prevalence similar to genes with reported duplications. The remaining genes had less than 10% prevalence in the population, although this level of prevalence is still striking.

We performed haplotype reconstruction on the transcriptome data to analyse the presence of CNV. Briefly, haplotype reconstruction leverages the process of VDJ recombination to examine which alleles of a particular IGHV gene are expressed with a particular IGHJ allele (in this study, as with many, IGHJ6, served as the anchor). As genes that undergo VDJ recombination exist on the chromosome, a haplotype can be reconstructed from examining which alleles exist together in certain VDJ combinations. We were able to resolve CNV on a chromosome level by examining the number of alleles assigned for each gene to a particular haplotype. As a result of higher IGHV gene coverage in the transcriptomic data, compared to the germline sequencing, we identified duplications in 18 genes (Fig. 3B). The duplications often consisted of the same IGHV gene alleles on both chromosomes. Furthermore, through haplotype analysis, we were able to resolve whether duplications occurred on a single chromosome or both chromosomes. Duplications on both chromosomes were present in 4/50 IGHV genes in this dataset. These indicate that there is inheritance of these duplications from both parents, suggesting a wider prevalence of IGHV gene duplication in our population. Due to the complexity innate to assigning alleles to particular genes, there may be overrepresentation of duplication for certain genes. An example of this is IGHV3-30 where allele *18 is identical to an allele of IGHV3-30–5. Indeed this allele is implicated for several participants as part of IGHV3-30 duplication and thus may not represent true duplication. Furthermore, no alleles were assigned to IGHV1-69D and IGHV2-70D which are duplicated genes and highly homologous to IGHV1-69 and IGHV2-70. Therefore, it is possible that some alleles assigned to the non- "D" genes are in fact alleles of those genes. However, this would not necessarily change the detection of duplication within this dataset. This also holds true for the genomic data produced in this study. This analysis is further complicated by the differences in expression between alleles of IGHJ6, as in the case of CAP088 and CAP256. The haplotyping process of RabHit was adjusted to account for this but may however be a limitation. Nevertheless, these data provide information on potential CNV events that can be validated by genomic techniques.

Unique VDJ counts were normalized as a proportion of the transcripts to control for variable VDJ counts between participants. These were then used to calculate a fold change in expression between haplotypes that did not have a duplication and those that did. With the reconstructed haplotypes from transcriptome data, we were able to connect the presence of a duplication with differences in expression and validate duplication observed previously in germline sequencing (Fig. 3C). While most of the genes demonstrate marginal fold changes as the result of gene duplication (> onefold change), there were six genes that demonstrate a much more distinct increase in gene expression when duplicated (IGHV1-3: 5.2 fold, IGHV3-9: 3.4 fold, IGHV5-51: 1.4 fold, IGHV1-69: 1.2 fold, IGHV2-70: 3.1 fold, IGHV3-73: 10.5 fold). This is consistent with what has been reported in other studies, where duplicated genes such as IGHV1-69 demonstrate biased usage (Avnir et al. 2016; Kenter et al. 2021; Pennell et al. 2023). The broad range (1.4–10.5 fold change) in expression levels amongst these genes suggests that duplication alone is not the only causative factor in enhanced expression indicating a more complex basis for this phenomenon. Some of these differences in expression could be explained by biased allele usage, which has been reported

in several genes (Ohlin 2021). This may particularly be the case for the duplicated haplotypes for IGHV1-3 which harbour an 01 allele which is expressed much more highly than the 02 alleles present in the haplotypes that lack duplication. While this is the only allele with reported usage biases that is present in the haplotypes of the dataset, it remains possible that as yet unknown allelic biases may be contributing to the observed differences in expression.

Discussion

Here we examined the genetic variation in both the IGHV sequences of a South African population, using multiple approaches. We observed significant levels of novel genetic diversity including in the leader peptide sequence, a functionally important portion of the gene that modulates expression. Importantly, many of these novel alleles were expressed as part of the antibody transcriptome, suggesting that they are utilized in the adaptive immune response. Furthermore, gene duplications were observed in the majority of the sequenced genes with some evidence of changes in expression levels as a result of duplication.

There are currently more than 300 alleles recorded in IMGT and other databases (Lefranc et al. 2015; Manso et al. 2022; Lefranc 2014; Lees et al. 2020). Through sequencing of 70 African donors, we have contributed an additional 32 novel IGHV alleles. We have also provided full-length sequences for 37 previously truncated alleles, allowing more confident inclusion of those sequences within human germline reference data. High levels of novel variation are consistent with several recent studies that examine the IGH locus (Watson and Breden 2012; Watson et al. 2013; Rodriguez et al. 2020; Corcoran et al. 2016; Gidoni et al. 2019; Hardt et al. 2023; Boyd et al. 1950). Other studies have examined African participants; however, our population yielded different alleles to those identified in largely Yoruban samples (Watson et al. 2013; Scheepers et al. 2015). It is probable that IG loci follow the population sub-structures identified more broadly in African genomes (Choudhury et al. 2020; Sengupta et al. 2020; Tau et al. 2017; May et al. 2013). Notably, however, the samples used in this study come from a single geographic population (KwaZulu-Natal, South Africa) and likely do not represent the broader genetic diversity across South Africa or the Southern African region. Further studies in cohorts across Southern Africa, and Africa more broadly, are likely to reveal novel genetic diversity.

Transcriptomic sequencing data derived from a subset of 8 of these individuals revealed further novel genetic diversity and explored linkage to gene expression. These results included confirmation that novel alleles discovered are expressed and contribute to a functional antibody response. Furthermore, the transcriptome sequencing showed better coverage across IGHV genes than the germline sequencing. This is likely the result of the sequencing method being more unbiased than the germline approach. The transcriptomic approach targeted highly conserved regions (framework 1 and the leader sequence) within the gene, whereas the germline sequencing used primers that target intergenic regions outside of the gene. Unknown polymorphisms in these intergenic regions may have impacted the ability of these primers to amplify all IGHV genes reliably. Our approach could be improved by a more complete understanding of the genetic diversity in intergenic regions in the IGH locus for Southern African individuals. Newer approaches, like hybrid capture described in Rodriguez et al. (2023), can provide complete characterization of both structural and allelic diversity in the IGH locus and could be applied in future studies.

There has been a large increase in the number of antibody transcriptomic studies being performed globally. Many of the data for these studies are housed in public databases such as iReceptor (Corrie et al. 2018) and VDJbase (Omer et al. 2020) and there are rigorous processes for the validation of novel alleles described from inference (Ohlin et al. 2019). This study indicates the utility of leveraging previously generated data to examine genetic diversity in IGH genes and creates a useful link to expression in the antibody transcripts. Meta-studies enable a deeper understanding of human genetic variation in the IG loci.

Full IGHV sequences were produced by amplicon sequencing, allowing for the examination of the leader sequence upstream of the variable gene. Similar to variable genes, substantial diversity was uncovered in the leader sequences. This has also been reported in the heavy chain (IGH) and light chain (IGK, IGL) loci of other populations (Mikocziova et al. 2020; Zhu et al. 2021; Mikocziova et al. 2021b; Huang et al. 2021). This study confirms that this phenomenon occurs in many of the same genes discussed in previous studies (IGHV1-18, IGHV1-58, IGHV2-26, IGHV4-4, IGHV4-61, IGHV5-51), inclusive of the mismatch variation described here. Furthermore, many variants identified in these studies are present in our dataset. For example, many of the IGHV4-4 alleles identified in this study demonstrate variation at IMGT position −31 in the leader between G and C, which is consistent with what has been reported in previous studies (Mikocziova et al. 2020; Zhu et al. 2021; Mikocziova et al. 2021b; Huang et al. 2021). These studies broadly focussed on European populations and here we described similar variation in African populations, suggesting that this phenomenon is common across different human populations. The uncovered mismatches in allele assignment between the leader and the variable regions may be the result of nucleotide variation that simply matches other alleles or could possibly represent a recombination event between alleles due to high sequence homology. It is probable that

References

Abdool Karim Q, Abdool Karim SS, Frohlich JA, Grobler AC, Baxter C, Mansoor LE et al (2010) Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. Science 329(5996):1168–1174

Avnir Y, Tallarico AS, Zhu Q, Bennett AS, Connelly G, Sheehan J et al (2014) Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. PLoS Pathog 10(5):e1004103

Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS et al (2016) IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci Rep 6(1):20842

Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD et al (1950) 2010 Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol Baltim Md 184(12):6986–92

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T et al (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 16(10):1784

Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T et al (2020) High-depth African genomes inform human migration and health. Nature 586(7831):741–748

Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA et al (2016) Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun 7:13642

Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E et al (2018) iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev 284(1):24–41

Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ et al (2014) Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature 509(7498):55–62

Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I et al (2019) Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nat Commun 10(1):628

Hardt U, Corcoran MM, Narang S, Malmström V, Padyukov L, Karlsson Hedestam GB (2023) Analysis of IGH allele content in a sample group of rheumatoid arthritis patients demonstrates unrevealed population heterogeneity. Front Immunol 14:1073414

Huang Y, Thörnqvist L, Ohlin M (2021) Computational inference, validation, and Analysis of 5'UTR-leader sequences of alleles of immunoglobulin heavy chain variable genes. Front Immunol 12:730105. https://doi.org/10.3389/fimmu.2021.730105

Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K et al (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe 16(1):105–114

Kenter AL, Watson CT, Spille JH (2021) Igh locus polymorphism may dictate topological chromatin conformation and V gene usage in the Ig repertoire. Front Immunol 12:682589

Lab H (2010) FASTX toolkit. Hannon Lab Cold Spring Harbor, NY

Lees W, Busse CE, Corcoran M, Ohlin M, Scheepers C, Matsen FA IV et al (2020) OGRDB: a reference database of inferred immune receptor genes. Nucleic Acids Res 48(D1):D964–D970

Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S et al (2015) IMGT, the international ImMunoGeneTics information system 25 years on. Nucleic Acids Res 43:D413-422

Lefranc MP (2014) Immunoglobulin and T cell receptor genes: IMGT(®) and the birth and rise of immunoinformatics. Front Immunol 5:22. https://doi.org/10.3389/fimmu.2014.00022

Lefranc M-P, Lefranc G (2001) The immunoglobulin factsBook, Academic Press, p 458

Manso T, Folch G, Giudicelli V, Jabado-Michaloud J, Kushwaha A, Nguefack Ngoune V et al (2022) IMGT® databases, related tools and web resources through three main axes of research and development. Nucleic Acids Res 50(D1):D1262–D1272

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1):10–2

May A, Hazelhurst S, Li Y, Norris SA, Govind N, Tikly M et al (2013) Genetic diversity in black South Africans from Soweto. BMC Genomics 14:644

Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G et al (2020) Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. Nucleic Acids Res 48(10):5499–5510

Mikocziova I, Greiff V, Sollid LM (2021a) Immunoglobulin germline gene variation and its impact on human disease. Genes Immun 22(4):205–217

Mikocziova I, Peres A, Gidoni M, Greiff V, Yaari G, Sollid LM (2021) Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes. iScience 103192. https://doi.org/10.1016/j.isci.2021.103192

Mor M, Werbner M, Alter J, Safra M, Chomsky E, Lee JC et al (2021) Multi-clonal SARS-CoV-2 neutralization by antibodies isolated from severe COVID-19 convalescent donors. PLoS Pathog 17(2):e1009165

Moyo T, Kitchin D, Moore PL (2020) Targeting the N332-supersite of the HIV-1 envelope for vaccine design. Expert Opin Ther Targets 24(6):499–509

Nielsen SCA, Yang F, Jackson KJL, Hoh RA, Röltgen K, Jean GH et al (2020) Human B cell clonal expansion and convergent antibody responses to SARS-CoV-2. Cell Host Microbe 28(4):516-525.e5

Ohlin M (2021) Poorly expressed alleles of several human immunoglobulin heavy chain variable genes are common in the human population. Front Immunol 11:603980. https://doi.org/10.3389/fimmu.2020.603980

Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D et al (2019) Inferred allelic variants of immunoglobulin receptor genes: a system for their evaluation, documentation, and naming. Front Immunol 10:435

Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT et al (2020) VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Res 48(D1):D1051–D1056

Pappas L, Foglierini M, Piccoli L, Kallewaard NL, Turrini F, Silacci C et al (2014) Rapid development of broadly influenza neutralizing antibodies through redundant mutations. Nature 516(7531):418–422

Pennell M, Rodriguez OL, Watson CT, Greiff V (2023) The evolutionary and functional significance of germline immunoglobulin gene variation. Trends Immunol 44(1):7–21

Peres A, Gidoni M, Polak P, Yaari G (2019) RAbHIT: R antibody haplotype inference tool. Bioinforma Oxf Engl 35(22):4840–4842

Richardson SI, Moore PL (2019) The antibody response in HIV-1-infected donors. Curr Opin HIV AIDS 14(4):233–239

Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M et al (2020) A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. Front Immunol 11:2136

Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT et al (2023) Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. Nat Commun 14(1):4419

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584

Safonova Y, Shin SB, Kramer L, Reecy J, Watson CT, Smith TPL et al (2022) Variations in antibody repertoires correlate with vaccine responses. Genome Res 32(4):791–804

Scheepers C, Shrestha RK, Lambson BE, Jackson KJL, Wright IA, Naicker D et al (2015) Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. J Immunol 194(9). https://doi.org/10.4049/jimmunol.1500118

Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K et al (2020) Genetic-substructure and complex demographic history of South African Bantu speakers. https://doi.org/10.1038/s41467-021-22207-y

Setliff I, McDonnell WJ, Raju N, Bombardi RG, Murji AA, Scheepers C et al (2018) Multi-donor longitudinal antibody repertoire sequencing reveals the existence of public antibody clonotypes in HIV-1 infection. Cell Host Microbe 23(6):845-854.e6

Soto C, Ofek G, Joyce MG, Zhang B, McKee K, Longo NS et al (2016) Developmental pathway of the MPER-directed HIV-1-neutralizing antibody 10E8. PLoS One 11(6):e0157409

Soto C, Finn JA, Willis JR, Day SB, Sinkovits RS, Jones T et al (2020) PyIR: a scalable wrapper for processing billions of immunoglobulin and T cell receptor sequences using IgBLAST. BMC Bioinformatics 21(1):314

Tau T, Wally A, Fanie TP, Ngono GL, Mpoloka SW, Davison S et al (2017) Genetic variation and population structure of Botswana populations as identified with AmpFLSTR Identifiler short tandem repeat (STR) loci. Sci Rep 7(1):6768

Vaccine genetics of IGHV1-2 VRC01-class broadly neutralizing antibody precursor naïve human B cells - PubMed [Internet] (2024) Available from: https://pubmed.ncbi.nlm.nih.gov/34489473/

van Loggerenberg F, Mlisana K, Williamson C, Auld SC, Morris L, Gray CM et al (2008) Establishing a cohort at high risk of HIV infection in South Africa: challenges and experiences of the CAPRISA 002 acute infection study Ugarte-Gil CA, editor. PLoS One 3(4):e1954

Vázquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M et al (2019) High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. Front Immunol 10:660

Walker LM, Phogat SK, Chan-Hui PY, Wagner D, Phung P, Goss JL et al (2009) Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science 326(5950):285–289

Watson CT, Breden F (2012) The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun 13(5):363–373

Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J et al (2013) Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Am J Hum Genet 92(4):530–546

Wheatley AK, Whittle JRR, Lingwood D, Kanekiyo M, Yassine HM, Ma SS et al (1950) 2015 H5N1 vaccine-elicited memory B cells are genetically constrained by the IGHV locus in the recognition of a neutralizing epitope in the hemagglutinin stem. J Immunol Baltim Md 195(2):602–10

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30(5):614–620

Zhou T, Lynch RM, Chen L, Acharya P, Wu X, Doria-Rose NA et al (2015) Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. Cell 161(6):1280–1292

Zhu Y, Yang X, Ma C, Tang H, Wang Q, Guan J et al (2021) Antibody upstream sequence diversity and its biological implications revealed by repertoire sequencing. J Genet Genomics Yi Chuan Xue Bao 48(10):936–945

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 2:

Identification of novel IgA and IgG antibody alleles in a South African population

Abstract

The Fc region of an antibody mediates interactions with cells of the immune system through binding to Fc receptors (FcRs). Genetic variation in the genes encoding the Fc region can impact these interactions, through altering binding kinetics. The Fc region is encoded by a set of IG (immunoglobulin) genes which also determine the isotype of the antibodies. Here we examine genetic diversity in IGHG and IGHA genes, which encode the IgA and IgG class of antibodies, in a South African population. We focus on these genes as they are most relevant to the study of viral infections. The genetic diversity of IGHG genes has been examined extensively, including in African populations, revealing a large number of alleles at variable demographic frequencies. In contrast, IGHA is less well understood with no genetic studies having been performed in the past 20 years. We performed targeted amplicon-based NGS in a South African population to examine if similar diversity is seen in IGHG and IGHA genes as was observed in the IGHV genes. For IGHA and IGHG we analysed sequences from 58 and 8 participants, respectively, identifying 65 total alleles, 83% (n=56) of which were determined to be novel. The largest number of novel alleles were found in IGHA1 (n=26) but novel alleles were also detected for IGHG1 (n=10), IGHG2 (n=7), IGHG3 (n=3) and IGH4 (n=9). Furthermore, within IGHA1, we characterised 9 new allotypes, including variation at position 411 (F>Y), a key residue in binding pIgR, facilitating antibody translocation across cellular membranes. Furthermore, mutations in IGHG that have been implicated in either enhancing or reducing several effector functions like ADCC were identified. Altogether we have described significant novel genetic diversity in IGHA and IGHG genes of African origin, which may have implications for the function of antibodies across different populations.

Introduction

Antibodies play a large role in the adaptive immune response. Several antibody classes exist across the various compartments of the human body and perform different functions. The fragment crystallizable (Fc) region determines how an antibody interacts with the broader immune system (1). The Fc region is encoded by the IGHC gene classes which also determine the isotype of the antibody: IGHM - IgM, IGHG - IgG, IGHE - IgE, IGHA - IgA, IGHD - IgD (2). These genes are all found within the IGH locus on chromosome 14. In this study, we focus on the IGHA and IGHG genes as they play the largest role in antiviral immunity (3). IgG is the most abundant isotype circulating in blood and is vital in engaging cells of the immune system (4). The IgA isotype is commonly found in the mucosae, in its dimeric form, where it is critical for mucosal immunity (5). The monomeric form of IgA can be found in human serum, however studies have demonstrated that IgA is most effective as a dimer (6).

Structurally, the Fc has three constant heavy domains (CH1-3) and a hinge (or several in the case of IgG3) between CH1 and CH2 (7). It is through interaction between the Fc region and various cellular receptors that an antibody enacts effector function (8). These include antibody-dependent cellular cytotoxicity (ADCC), antibody-dependent cellular phagocytosis (ADCP), antibody-dependent complement deposition (ADCD), and trogocytosis. The cellular receptors that bind the Fc region are called Fc receptors or FcRs, often binding through interactions with CH2 (9). There are seven FcRs with variable affinities for different antibody isotypes that are expressed on a variety of cells in the body.

The IGH locus, which contains all of the constant region genes, is one of the most polymorphic and structurally diverse regions of the human genome (10). The most allelically diverse gene is IGHG3 which has 29 recorded alleles in IMGT, the gold standard immunogenetic reference database (11). The next largest are IGHG1 and IGHG2 which have 14 and 17 alleles, respectively. Finally, IGHG4 has only eight alleles but interestingly exhibits gene duplication, which is more commonly found in IGHV genes (12,13). These reference data have been produced in large recent studies from different populations in geographic locations across the world such as Africa and Brazil (13–15). IGHA, by contrast, has very few known reference alleles (only four for IGHA1 and three for IGHA2) (11). This is largely due to there being no known large-scale genetic studies on IGHA genetic diversity.

Protein diversity in IGHA and IGHG genes has previously been described through serological methods (16). The various protein variants are termed allotypes, where one allotype

represents a specific haplotype of amino acid mutations. In recent IGHG genomic studies, amino acid mutations that result from genetic variation have given rise to newly described allotypes (13–15). Currently, there are 46 described allotypes for IGHG genes (IGHG1=12, IGHG2=12, IGHG3=20, IGHG4=2) (13,17). While there is no described allotypic variation for IGHA1, two allotypes have been identified for IGHA2 (17). For ease of reading, we will also describe novel protein variants as allotypic variations to easily discriminate between specific amino acid residue changes and a haplotype of amino acid changes.

There are functional implications for variation in the protein sequence of the Fc region as demonstrated in a series of papers that have been published examining the effect of mutations on key elements of the IgA Fc (5, 18,19). Many of these lead to enhanced binding of FcαRI and consequently increase effector function (18,19). For the IGHG genes, both naturally occurring and induced mutations have been examined for their impact on Fc function. Allelic variation in CH2 has been demonstrated to enhance binding to FcγRIIIa and other FcRs leading to enhanced effector functions such as ADCC (20–27). Fc mutations can also affect other aspects of antibody function. CH3 mutations have been identified that create bispecificity in IgG antibodies, which either occur naturally or can be engineered (22,26,28–30). Bispecificity is created when antibodies with different paratopes are combined to create a new antibody that has the specificities of both. These have been useful in cancer immunotherapy but have also been developed to make antibodies that bind pathogens at multiple sites or from multiple strains.

This study aimed primarily to address the relative lack of information on IGHA genetic diversity, specifically within a South African population, and to contribute valuable new sequences. We also examined if the IGHG genetic diversity observed in this population is similar to what has been reported in other studies where there is considerably more data. We further aimed to determine through structural analysis if the genetic variants that we uncovered could have potential functional implications.

Methods

Participant samples and ethics

The same CAPRISA samples, from cohorts 002 and 004, from the paper in Chapter 1 were utilised here, under the same ethics clearance. This is also detailed in Samples and Ethics.

Germline IGHA library preparation and sequencing

Genomic DNA was extracted from participant PBMCs using the Wizard® Genomic DNA Purification Kit (Promega; Madison, WI USA). To generate IGHA genomic datasets, the DNA first underwent a gene-specific PCR amplification. In-house primers, initially designed for Sanger sequencing, were adapted for PacBio library preparation with the addition of an adapter section on the 5' end of each primer.

Forward primer:

5’-GCAGTCGAACATGTAGCTGACTCAGGTCACGGGCCGCGTCCTCACAGTGCATTC-3’

Reverse primer:

5’-TGGATCACTTGTGCAAGCATCACATCGTAGCAGACAGGCGCCGGATGGAAGCG-3’

Due to high sequence homology between IGHA1 and IGHA2 and evidence from prior Sanger sequencing, a single primer pair was used to amplify both genes. However, the primer pair was strongly biased towards the amplification of IGHA1, as only a single IGHA2 sequence was recovered during analysis. Each PCR reaction contained 14.75 µL of molecular grade water, 5 µL of 5x Platinum Superfi Buffer, 0.5 µL of a dNTP mix (10 uM each), 1.25 µL of forward primer (10 uM), 1.25 µL of reverse primer (10 uM), and 0.25 µL of Platinum SuperFi II Polymerase (Thermofisher Scientific; Waltham, MA USA) (2.5 U/µL). To this mix 2 µL of DNA (10 ng/µL) was added. The cycling conditions were as follows:

Initial Denaturation	98ºC	30s	
Denaturation	98ºC	10s	25 Cycles
Annealing	60ºC	10s	
Extension	72ºC	1m	
Final Extension	72ºC	5m	

Each sample was run in duplicate to aid in the elimination of erroneous reads during data analysis. Once this initial amplification was complete the amplicons were purified, eliminating remaining PCR reagents, with Ampure XP beads (Beckman Coulter; Brea, CA, USA). These beads bind DNA and were added at a 1:0.75 ratio, to ensure the removal of small DNA fragments (<200 bp).

The cleaned PCR product was evaluated with a Qubit Flex (Thermofisher Scientific; Waltham, MA USA), and positive amplification demonstrated a high concentration of DNA. After this QC step, the duplicates were mixed. To each participant sample, a unique PacBio barcode was added and another PCR was performed with the same Platinum SuperFi II Kit, with the following cycling conditions:

Initial Denaturation	98ºC	30s	
Denaturation	98ºC	10s	15 Cycles
Annealing	60ºC	10s	
Extension	72ºC	1m	
Final Extension	72ºC	5m	

The PCR was cleaned again using magnetic bead purification and submitted to the NICD sequencing core for Next Generation Sequencing. Before sequencing on a PacBio Sequel II, the samples were evaluated on both Qubit and Agilent Tapestation (Agilen; Santa Clara, CA, USA) to ensure that there was sufficient concentration and correct amplicon size. The library preparation was done using the SMRTbell Template Prep Kit v. 1.0 (Pacific Biosciences, Menlo Park, CA, USA). Sequencing was carried out on PacBio® Sequel system using SMRT cell 1M with a movie time of 600 min, pre-extension time of 90 min using Sequel Binding Kit v. 2.1, Sequel Sequencing Plate v. 2.1 and loading was done by diffusion. The resulting subreads were demultiplexed and the circular consensus sequences (CCSs) generated using the “ccs” command (0.9999 minimum predicted accuracy and 3 minimum passes) on SMRT Link v. 6.0 (Pacific Biosciences).

Previously generated data was used to examine IGHG diversity, which was produced in the same way as in this study with different primers (see next page). The original study was a pilot testing the utility of PacBio to sequence IGHG3. However, due to high sequence homology between the different IGHG genes the sequencing produced sequences for all four genes. In this study, we re-examine this data to categorise variation in IGHG genes for eight of the CAPRISA participants.

Forward Primer:

5’-GCAGTCGAACATGTAGCTGACTCAGGTCACGCTCTGTCCCACACCGCAGTCA-3’

Reverse primer:

5’-TGGATCACTTGTGCAAGCATCACATCGTAGGGGGTACGTGCCAAGCATCCTC-3’

Data analysis

The data output of a PacBio sequencing run is in the form of CCS reads. The following description of the data process and analysis was performed on all of the IGHG and IGHA genes. Low-quality reads were discarded after being processed by VSEARCH (Q score < 30)(31). The reads were then oriented correctly as PacBio reads can be in either strand, correct 5' to 3' orientation or in the reverse complement. This was achieved using a custom Julia script (all scripts were developed for this study and are hosted at: https://github.com/NICD-AIRU/AMarsden_IGHC). The same script also removed primer and adapter sequences before deduplicating the sequences. Reads with a depth of less than ten were discarded. The filtered sequences were then passed through a local BLAST to determine the identity of the sequences. The BLAST database contained full-length reference IGHC sequences from IMGT. These results were then parsed through a separate Julia script. The purpose of this script was to eliminate sequences with low identity to functional IGHA and IGHG genes, as pseudogenes and chimeras are possible within this dataset. Any sequence with more than ten mismatches from the closest reference sequence was discarded, as these were deemed likely chimeras due to high sequence homology between functional and non-functional IGHA genes. However, it is possible that this meant eliminating complex novel sequences with large mismatch but given the large number of novel alleles identified in this study, that is not likely. Furthermore, novel IGHA alleles from a newly published study also vary by less than ten nucleotides, thus reinforcing this filtering criteria (32). The sequences were then compiled into separate gene-specific FASTA files.

The gene-specific FASTA files were then processed by a custom Julia analysis script, which was run per gene. In brief, the script limited sequences to the top two most abundant sequences per individual. This was performed under the assumption that no gene duplication occurs, which has yet to be reported for IGHA and IGHG1-3 genes. While duplication has been reported for IGHG4, it does not seem to be a common phenomenon in African populations (32). All of the sequences were then collapsed from all participants to identify unique alleles. Once alleles were assigned to sequences, genotypes were described for all

participants. Each allele was then examined by extracting the CH regions. These were compared to reference sequences to identify mutations on both a nucleotide and amino acid level. Requisite FASTA files were produced for all steps as appropriate.

To represent nucleotide diversity, unique alleles for each gene were loaded into Geneious. From this, the variant frequency was determined (in comparison to the most abundant match to reference allele from IMGT) and alignments were visualised utilising the Highlighter tool from LANL (https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html). The amino acid diversity was also shown by demonstrating where changes lay on a crystal structure. These were obtained from PDB (IGHA1:1IGA, IGHG1:6IFJ, IGHG2:4HAG, IGHG3:6D58, IGHG4:5W5N) and PyMol was utilised for visualisation.

Results

Overview of IGHA and IGHG sequencing

In this study, we utilised genomic NGS to characterise IGHA and IGHG genetic diversity in an African population. The genomic sequencing for IGHA was performed on 58 participants yielding 27 unique allele sequences. Only a single IGHA2 sequence was obtained and matched the reference 01 allele. For IGHG, genomic sequencing data from an unpublished study of eight participants was used to determine if the variants we uncovered are similar to those previously reported in African populations. Novel alleles were called when nucleotide differences were observed compared to reference sequence data in IMGT, the immunogenetic reference database.

The greatest number of alleles, 27, were identified in IGHA1 sequences (Figure 1A). There were fewer alleles detected for IGHG overall, due to the smaller sample size. For IGHG1, 11 alleles were detected, 9 alleles were observed in IGHG2, 8 alleles were found for IGHG3, and 11 alleles for IGHG4.

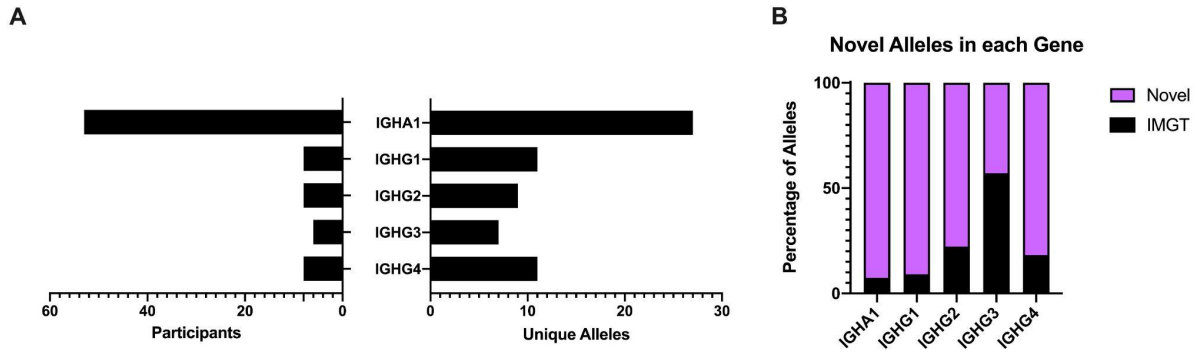


Figure 1: Description of IGHA and IGHG germline sequencing. A: The left-hand bar graph shows the number of participants that were sequenced for each IGHC gene. The right-hand bar graph shows the number of unique alleles identified for each IGHC gene. B: This demonstrates the proportion of unique alleles that were designated as novel (in purple) or matched IMGT reference data (in black).

Amongst these unique alleles, many novel alleles were identified (Figure 1B). In IGHA1, 26 of the 27 alleles were novel. Amongst the IGHG alleles, 29 alleles were novel (IGHG1:10,

IGHG2:7, IGHG3:3, IGHG4:9). In total, this makes 56 novel alleles identified in the study population.

IGHA1 nucleotide and amino acid variation

Next, we investigated the location and frequency of the nucleotide diversity in the 27 IGHA allele sequences produced by NGS from 58 CAPRISA participants compared to a reference sequence (Figure 2).

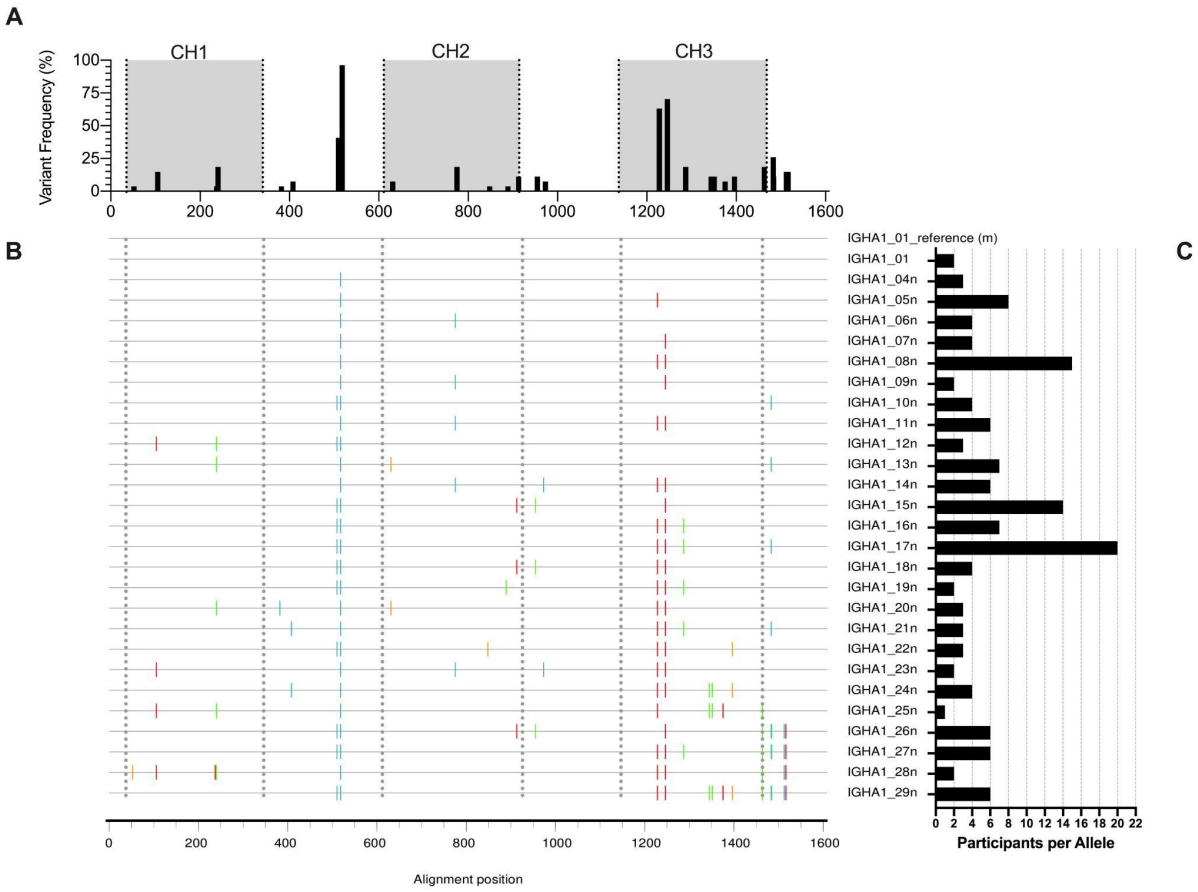


Figure 2: Distribution of nucleotide diversity in IGHA. A: Shows the frequency and position of SNPs across the whole IGHA1 gene. B: A highlighter plot showing the position of observed SNPs within the alleles. Novel alleles are designated with an 'n' in the allele name. C: Shows the number of participants for which each allele was observed.

In total, there were 27 positions across the whole ~1.5 kbp gene where variation was observed in the dataset. All of these were single nucleotide polymorphisms (SNPs), which were often present across multiple alleles. These SNPs occurred at different frequencies, ranging from

3.7% to 96.8% (Figure 2A). When compared to the reference IGHA1 alleles (IGHA1*01, IGHA1*02, IGHA1*04, IGHA1*06), all sequences, except for one which matched IGHA1*01, differed from the reference alleles. The novel alleles contained a range of SNPs (from 1 -> 10). Of the 27 SNPs, 17 occur within the CH regions of IGHA1. The most diversity was observed in CH3, where eight variants were uncovered. The other regions, CH1 and CH2, had fewer with four and five respectively. Interestingly, there were seven alleles where the changes occurring in the CH3 regions were completely identical to the CH3 region observed in IGHA2. We did not consider these as chimeras because there was sufficient read depth (>10 reads), indicating these may be instances of gene conversion which commonly occurs in IG constant genes (13,14).

Next, the amino acid and allotypic variation that results from the novel genetic diversity was examined (Figure 3A, Table 1). We have identified several novel allotypes, designated with an 'n'. In total, there were nine novel allotypes revealed. These novel allotypes result from seven different amino acid changes that occurred across the three CH regions, with 4 of 7 found within CH3, consistent with the nucleotide changes observed. The most prevalent of these novel allotypes (*02n) contains a single amino acid change, R392H, in CH3 and occurs in 39.7% (n=23) of participants. The second most frequent allotype (*03n) contained a single change in CH1, E295D, which was present in 20.7% (n=12) of participants The remaining allotypes (*04n - *10n) have a variety of mutations in the CH3, having a combination of R392H, F411Y, D428E, L451M mutations.

We further examined the location of shared amino acid mutations in the structure of an IgA1 antibody (Figure 3C). One mutation (F411Y), which occurs in three of the novel allotypes uncovered, alters a position within the CH3 region. This position is critical for the binding of pIgR (Polymeric Immunoglobulin Receptor), although it is unclear the effect this mutation may have on pIgR binding (33,34).

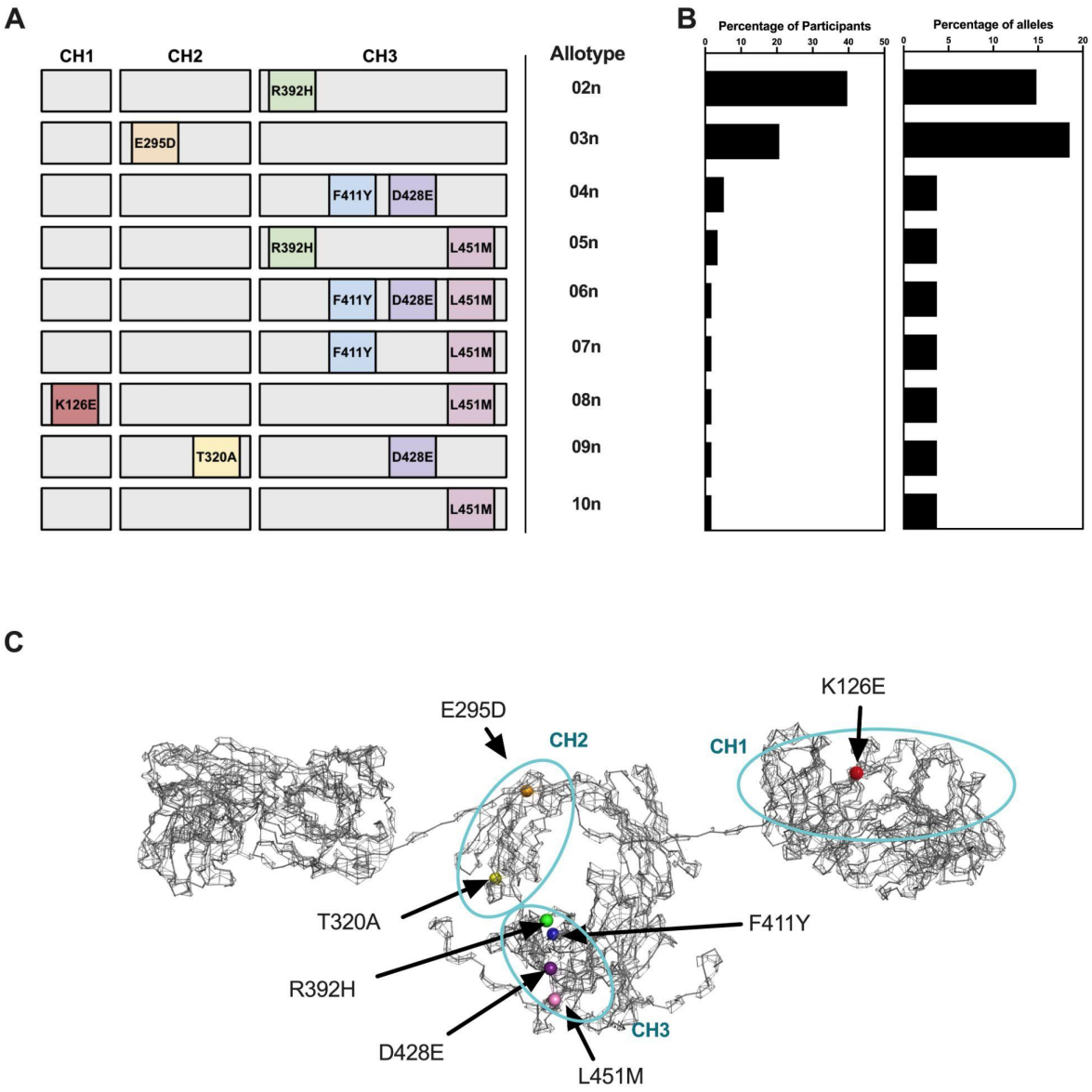


Figure 3: Novel allotypic variation in IGHA1. A: Graphical display of the different allotypes identified in the dataset, with each change noted in its appropriate region using Burr numbering (the standard for numbering amino acid changes in IGHA). B: The proportion of participants with each allotype and the proportion of unique alleles that translate into that protein variant. C: A protein model demonstrating the location of each amino acid change in the structure of an IgA1 antibody. The CH regions (CH1-CH3) are circled and labelled in blue.

Table 1: Alleles associated with IGHA1 allotypes

Allotype	Alleles
01	01, 05n, 12n, 15n, 18n, 08n, 20n, 13n, 04n, 10n, 07n
02n	17n, 19n, 21n, 16n
03n	11n, 14n, 06n, 09n, 23n
04n	24n
05n	27n
06n	29n
07n	25n
08n	28n
09n	22n
10n	26n

IGHG nucleotide and amino acid variation

The IGHG sequences from the eight CAPRISA participants were analysed in the same way as the IGHA sequences, identifying the location and frequency of SNPs across the genes. The number of SNPs across the four IGHG varies although they, like the IGHA sequences, lay predominantly in the CH regions of the genes, with others occurring in the intron mostly, with few in the hinge regions (Figures 4 and 5).

For IGHG1, 10/11 of the sequences were closely matched with the *05 allele, except for one which was similar to *02. There were 17 SNPs that occurred in IGHG1 with 11 SNPs within the CH regions (CH1:4, CH2:1, CH3:6) (Figure 4A-C). While most of these SNPs occurred at low frequencies amongst the alleles (<10%), 4 positions exhibited more prevalent mutations (45-72%). The IGHG2 sequences were split between identity to *02 and *06 reference alleles. There were fewer SNPs identified in IGHG2 (13 total) with eight of these variants present in CH regions (CH1:2, CH2:1, CH3:5) (Figure 4D-F). Many of the SNPs identified in IGHG2 were present at higher levels with eight being above 10%.

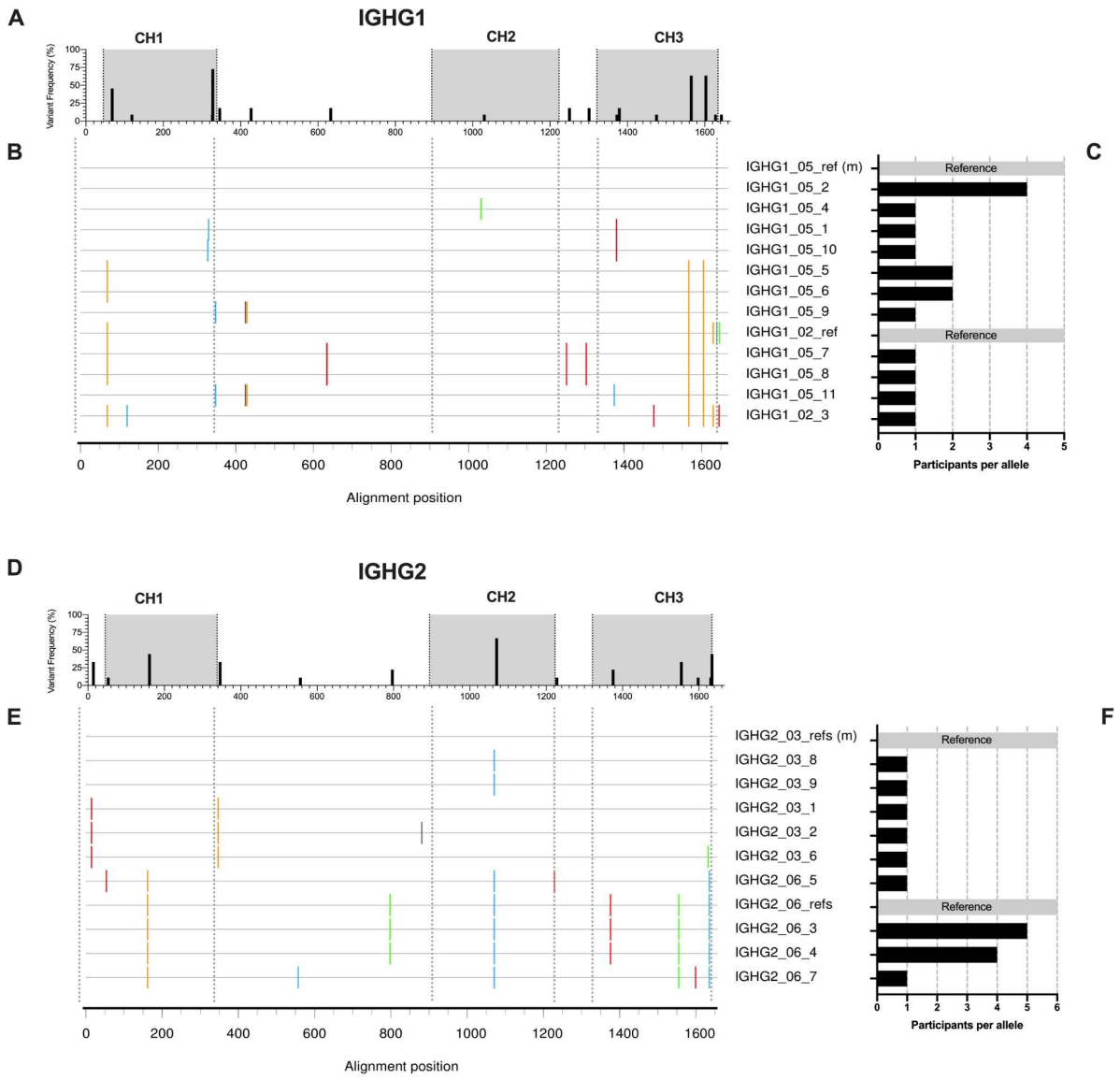


Figure 4: Nucleotide diversity in IGHG1 and IGHG2 genes. B, E: Each highlighter plot shows the SNPs observed in the dataset in comparison to IMGT reference sequences. A, D: Above each alignment, a graph represents the frequency at which particular variants are observed within the dataset. C, F: the right of each graph is a graph showing the number of participants each allele was observed in. A-C: shows IGHG1 and D-F: shows IGHG2

IGHG3 (Figure 4A-C) is the longest of the IGHG genes due to hinge duplication, which can vary between three to four hinges. This can be seen within the sample data where participant alleles that closely match IGHG3*03 contain a large gap between CH1 and CH2, in contrast to other IGHG3 alleles. This is the result of these sequences having two hinges as opposed to three (Figure 5A). The other alleles closely matched the reference *10 allele. Overall, 17 SNPs were observed in IGHG3 with most of these occurring in CH3. These SNPs were

observed at relatively high (14-70%) levels, although the number of participants is lower than other genes. In IGHG4 (Figure 4D-F), the alleles closely matched the reference *04 allele with more SNPs than in other genes (n=23), however, these were overall less frequent across the dataset (with 21 SNPs exhibiting frequencies between 9-18%). The vast majority of these (17 SNPs) lay in the CH regions, predominantly in CH3 (10 SNPs).

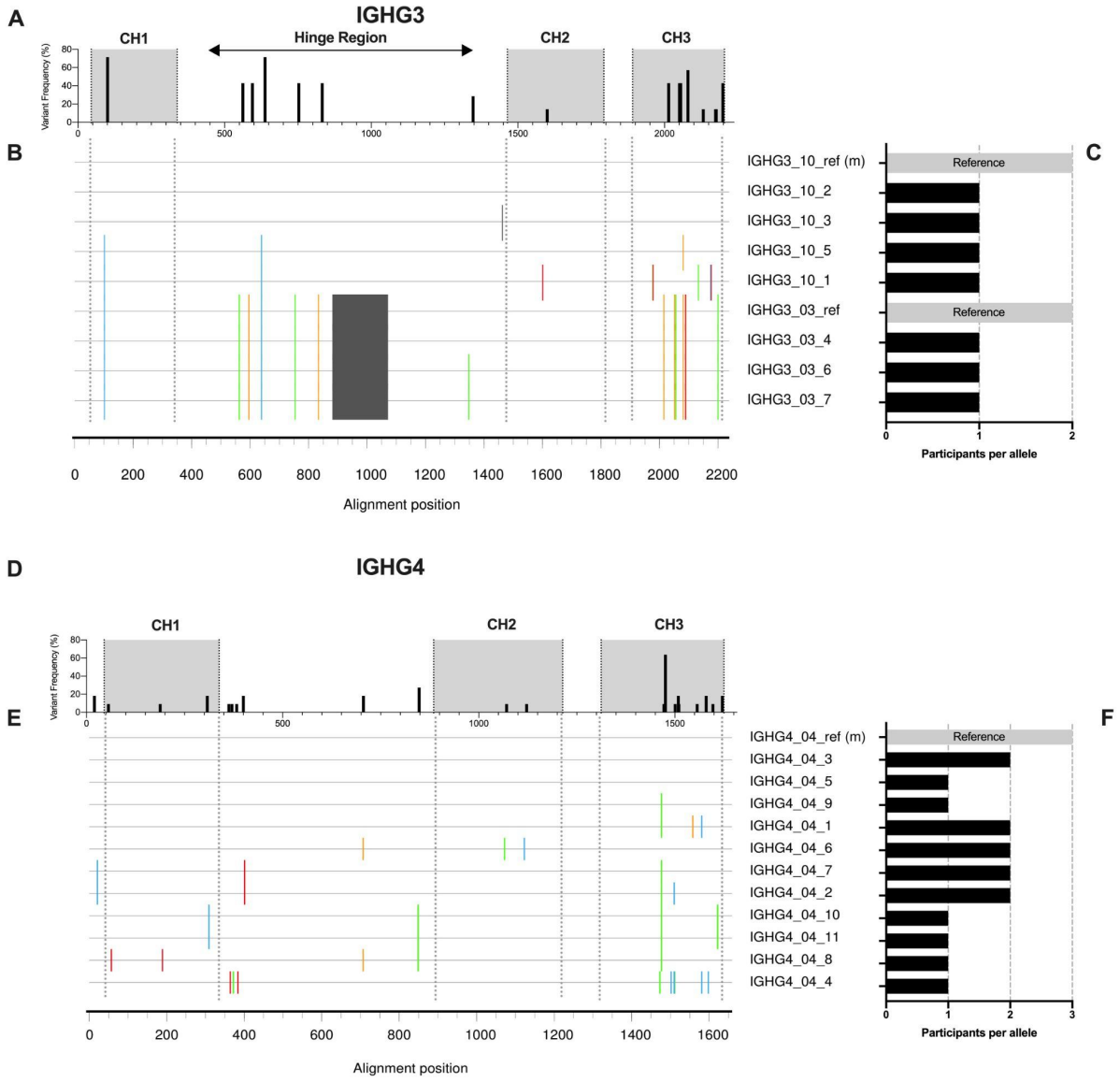


Figure 5: Nucleotide diversity in IGHG3 and IGHG4 genes. B, E: Each highlighter plot shows the SNPs observed in the dataset in comparison to IMGT reference sequences. A, D: Above each alignment, a graph represents the frequency at which particular variants are observed within the dataset. C, F: the right of each graph is a graph showing the number of participants each allele was observed in. A-C: shows IGHG3 and D-F: shows IGHG4

Next, we examined how this nucleotide diversity may affect the protein of an IgG antibody, matching the allotypes to previously reported allotypic diversity (Figure 6, Table 2). Despite the allelic diversity observed within IGHG1, only one allotype (the reference *01 allotype) was observed. In IGHG2 (In red, Figure 6), there were two allotypes observed, *06 and the reference *01. The variant allotype observed (*06), consisted of an A378S change accounting for 20% (n=2) of IGHG2 alleles. This position is significant as the change to a serine residue at this position has been demonstrated to adversely affect complement deposition (35).

IGHG3 (blue in Figure 6) had three different allotypes that differed from the reference *01 allotype in our population. The first allotype (*01m) contained a Q419E mutation, which has previously been described in a South African population from the CAPRISA cohort and has been associated with increased binding of FcγRIII and FcγRIIb, leading to enhanced ADCC responses (15). The allotype occurred in one participant. The next allotype (*14) contained two amino acid changes, P291L and S384, the first of which has been determined to reduce ADCC responses when a leucine residue occurs at position 291 in the CH2 region (36). The final IGHG3 allotype (*03) had both the four amino acid changes and was the most prevalent - occurring in half the participants. The allotype was defined as containing: M397V, K409R, Q419E, and I422V. Both positions 397 and 409 have been altered in another study to facilitate the swapping of Fab arms, creating bispecific antibodies (37).

The last gene, IGHG4 (in green in Figure 6), was determined to have three novel allotypes, as well as the reference *01 allotype. These novel allotypes were less frequently observed in the dataset than other IGHG variants, occurring in either one or two individuals and all resulting from a single allele. The first identified allotype (*01n) contained an L445P mutation towards the end of CH3. The next allotype (*02n) had two amino acid changes in CH2, V308I and N325. However, for the last allotype (*03n), there were similar mutations to what was observed in IGHG3: R409K, E419Q, and V422I. It is important to note that these are the opposite of what was detected in IGHG3, e.g. R409K -> K409R.

When we examined where the mutations were positioned in the structure of an antibody, all the changes existed within the CH2 and CH3 regions (Figure 6 C-E). Furthermore, many of the mutations observed on the surface of the protein increasing the likelihood that they affect interactions between the Fc and various receptors and proteins.

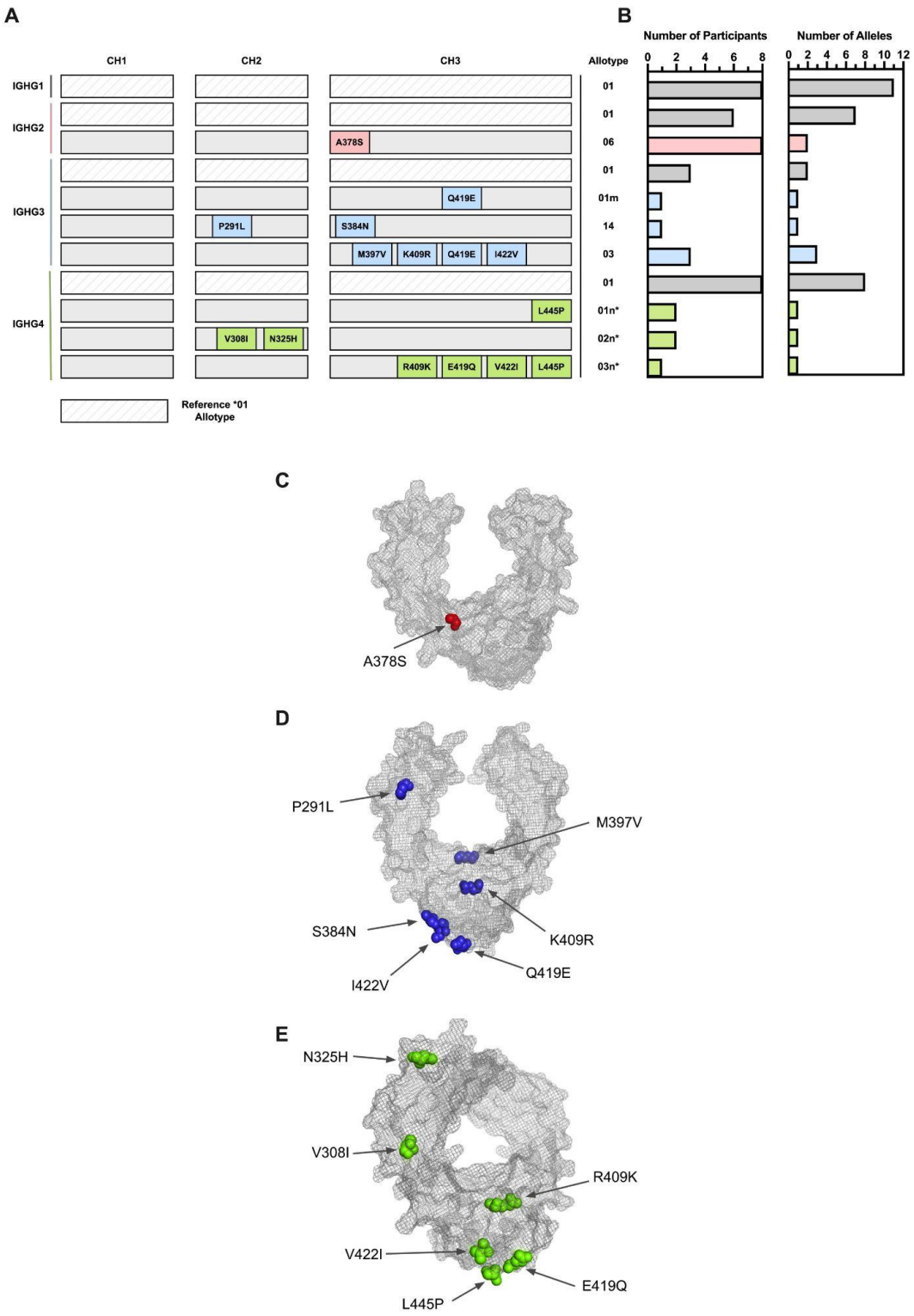


Figure 6: Allotypes in IGHG. A: Allotypes, as they differ from the 01 allotype, identified in the dataset, with each change noted in its appropriate region using Eu numbering. The allotype name is to the right, with the 01m allotype identified from prior studies and the 'n' IGHG4 allotypes representing novel allotypes identified in this study. B: The left bar shows the percentage of participants in which the variant is found, and the right graph demonstrates the percentage of unique alleles that translate into that protein variant. The genes are differentiated by colour, with IGHG2 in found,GHG3 in blue, and IGHG4

in green. IGHG1 demonstrated no protein variation in the data set. C-E: Protein variation in the structure of IGHG Fc region (CH2-CH3). All three are protein models of the Fc portion of IgG antibodies. Amino acid changes are indicated by spheres and labeled accordingly. C represents IGHG2, D represents IGHG3 and E represents IGHG4.

Table 2: Alleles associated with IGHG allotypes

Gene	Allotype	Alleles
IGHG1	01	IGHG1*05_1, IGHG1*05_2, IGHG1*02_3, IGHG1*05_4, IGHG1*05_5, IGHG1*05_6, IGHG1*05_7, IGHG1*05_8, IGHG1*05_9, IGHG1*05_10, IGHG1*05_11
IGHG2	01	IGHG2*03_1, IGHG2*03_2, IGHG2*06_5, IGHG2*03_6, IGHG2*06_7, IGHG2*03_8, IGHG2*03_9
	06	IGHG2*06_3, IGHG2*06_4
IGHG3	01	IGHG3*10_2, IGHG3*10_3
	01m	IGHG3*10_5
	14	IGHG3*10_1
	03	IGHG3*03_4, IGHG3*03_6, IGHG3*03_7
IGHG4	01	IGHG4*04_2, IGHG4*04_3, IGHG4*04_5, IGHG4*04_7, IGHG4*04_8, IGHG4*04_9, IGHG4*04_10, IGHG4*04_11
	01n	IGHG4*04_1
	02n	IGHG4*04_6
	03n	IGHG4*04_4

Discussion

In this study, we have uncovered novel IGHG and IGHA alleles by sequencing the antibody genes of a South African population group. Furthermore, we have demonstrated that some of this novel genetic diversity contributes to alterations in the amino acid sequences of IgG and IgA antibodies. These changes may have implications for how FcRs bind to different protein variants which can impact how antibodies induce a variety of effector functions with consequences for dealing with infectious agents.

Overall, we describe 56 novel IGHA and IGHG alleles. Although IMGT contains various IGHA and IGHG alleles that are truncated, none of the novel alleles described in this study represent extensions of these truncated alleles and are entirely novel. A majority of the novel alleles were found in IGHA1, likely due to the greater sample size and the fact that IGHA is studied far less than IGHG. The lack of IGHA2 sequences is likely the result of primer biases in the original amplification performed on the samples and also may potentially be impacted by undescribed nucleotide variation in the primer binding regions. Although we sequenced fewer participants for IGHG, we were able to identify a series of novel alleles across all four IGHG genes. Some of these sequences may be chimeras due to the high sequence homology between the genes, however, it is possible that these represent gene conversion events which have also been reported in other studies (13,14). IGHG3 produced fewer sequences than the other IGHG genes, which is possibly the result of primer biases that favoured IGHG1, IGHG2 and IGHG4. This could be remedied with an alternative approach to library preparation taking into account hinge length variation. While SNPs were identified in the other parts of the gene, such as the intron and the hinge, their relevance is unclear.

A previous study by Bashirova *et. al* examined 94 participants from the same geographic locale (Kwa-Zulu Natal, South Africa) as the samples used for this study, albeit from different cohorts, amongst other populations (13). Bashirova *et. al* utilised Sanger sequencing of the CH regions of the genes to characterise allotypic diversity and uncovered novel variants, which we have used as the basis of naming the allotypes detected in our study. The study found that the populations demonstrated inter-population differentiation in IGHG1, IGHG2 and IGHG3, where certain alleles were present at different frequencies and in different haplotypes across the populations. The South African populations had haplotypes that differed from European samples and were also distinct from African-American samples. For example, a haplotype consisting of IGHG1*02, IGHG2*06, and IGHG3*01 was prevalent in South African samples but altogether missing in European samples. The study also examines the prevalence of

different IGHG allotypes within the South African population. In both this, and the previous study the IGHG1 reference (*01) allotype is the most prevalent allotype within a South African population. However, for IGHG2 the most prevalent allotypes detected in Bashirova *et. al* were *03 and *06, while our data demonstrates a high prevalence of *01 and *06. This discrepancy is likely due to the sample size differences between the two studies. Lastly, when comparing IGHG3, Bashirova *et. al* also reported that the most prevalent allotypes are *01 and *03, which is in agreement with our dataset. Therefore, despite differences in technique and sample size there is a strong overlap in the distribution of allotypes between our dataset and those reported in Bashirova *et. al*.

A sequencing method has been published which allows for resolution of an entire antibody transcript including the full Fc region, termed FLAIRR-Seq (full-length adaptive immune receptor repertoire sequencing, 38). Through this method the original study also uncovered novel IGHG alleles. There was overlap between the alleles identified here and those described in the study. Overall, there were seven IGHG4, two IGHG1 and one IGHG2 novel alleles from our dataset that matched alleles described using FLAIRR-Seq. The approach provides a promising avenue for combining bulk VDJ sequencing with Fc variant discovery to create a powerful dataset to examine the relationship between IG germline diversity and the functional antibody response. Future studies in African populations can utilise this methodology to comprehensively genotype a vast portion of the IGH locus.

Furthermore, whole IGH locus sequencing has also uncovered novel genetic diversity across a breadth of demographics, both polymorphic and structural (33). The African participants that were sequenced in the study demonstrated higher levels of novel allelic variation than other populations, which reinforces the large number of novel alleles described here. There are alleles described here that are also observed in this whole locus study (IGHA1: 9, IGHG1:3, IGHG2:1, IGHG3:4 and IGHG4:3). This overlap indicates that the novel variation uncovered in this study is not wholly contained within the specific South African population sampled. Furthermore, the large number of novel alleles both described here and in the study that aren't shared between different populations reveals that variation in IG constant genes is much deeper and more widespread than initially suggested.

We examined how genetic diversity in IGHG and IGHA genes created alterations in the protein sequence of the Fc region of antibodies (summarised in Table 3). In the case of IGHA1, we described 9 novel allotypes that differ from the reference IGHA1 sequence. The potential consequences of most of these amino acid changes are unknown. However, previous mutagenesis experiments have implicated an F residue at position 411 in CH3 as affecting the translocation of dimeric IgA through binding to pIgR (32,39,40). Abundant in the

gastrointestinal tract, pIgR binds to the CH3 region of dimeric IgA to facilitate movement of the antibody across the outermost epithelial membrane of the mucosae, allowing for antibody mediated protection at the mucosal surface. Whether the F411Y mutation identified here will have a similar impact will require functional characterisation. Such investigations could lead to the identification of additional mutations that improve functionality leading to broader efficacy of antibody therapy.

Although fewer of the allotypes that were uncovered for the IGHG genes were novel, we observed amino acid changes that have described effects for an IgG antibody (summarised in Table 3). There were three novel allotypes identified for IGHG4. In IGHG, three positions (291, 378, 419) lead to an increase or decrease in effector function, two through ADCC and one via CDC. Participants who have these mutations may therefore have an impact on their immune response against an infection or vaccination. The remaining two mutations are known to create bispecificity (37).

Table 3: Summary of mutations in IGHG and IGHA sequences from this study and their potential effects

Gene	Position	Participants	Allotypes	Binding	Effect
IGHA	411	5	3	pIgR	Translocation of IgA across epithelial membranes
IGHG	291	1	1	FcγRIIIA	Reduction of ADCC response if Leucine at position
	378	8	1	C1q	Decrease in CDC activation if Serine at position
	397	3	1	-	Creation of bispecificity through Fab arm swapping
	409	5	2		
	419	5	2	FcγRIII and FcγRIIb	Enhancement of ADCC if Glutamine at position

There are several ways to examine the functional implications of diversity in the Fc region. A previous study has demonstrated that Fc effector function assays can be used to examine how antibodies with Fc region variation lead to alterations in the recruitment of ADCC (15). This approach can be utilised to understand how the mutations described in this study can affect how antibodies with these mutations can alter effector function. Furthermore, binding assays can also be utilised to examine the impact of Fc mutations on binding to FcRs, which will indicate the impact the mutations have on effector function. It is possible that, in addition to impacting effector function, the mutations we have uncovered may lead to differing serum levels between the various protein variants through altered class-switching and receptor binding (41). Fluorescence microscopy is a common method to study the translocation of proteins in tissues and cells and has been used to examine the role of IgA antibodies in HIV infection of mucosal surfaces previously (42,43). This is an approach that can be used to categorise how altered pIgR binding in IgA affects the movement of IgA across mucosal membranes.

It is important to note that the genetic diversity described in this study comes from a particular South African population. Performing sequencing in other South African and broadly African populations will likely yield more novel diversity. Future studies understanding how mutations in the Fc region affect the function of an antibody may enable better understanding of how different populations mount immune responses against pathogens. Moreover, these mutations may be useful for the production of therapeutic mAbs, allowing for the creation of more effective interventions.

References

1. Prinz W, van Dixhoorn MS, Mayne E, Wadee A. Immunology. In: Molecular Medicine for Clinicians. 2009. p. 289–310.

2. Lefranc MP, Lefranc G. The Immunoglobulin Factsbook. 2001. 72–239 p.

3. Muramatsu M, Yoshida R, Yokoyama A, Miyamoto H, Kajihara M, Maruyama J, et al. Comparison of antiviral activity between IgA and IgG specific to influenza virus hemagglutinin: increased potential of IgA for heterosubtypic immunity. PloS One. 2014;9(1):e85582.

4. Leusen JHW, Nimmerjahn F. The Role of IgG in Immune Responses. In: Nimmerjahn F, editor. Molecular and Cellular Mechanisms of Antibody Activity. New York, NY: Springer New York; 2013. p. 85–112.

5. Woof JM, Russell MW. Structure and function relationships in IgA. Mucosal Immunol. 2011 Nov;4(6):590–7.

6. de Sousa-Pereira P, Woof JM. IgA: Structure, Function, and Developability. Antibodies. 2019 Dec 5;8(4):57.

7. Schroeder HW, Cavacini L, Cavacini L. Structure and function of immunoglobulins. J Allergy Clin Immunol. 2010 Feb;125(2 Suppl 2):S41-52.

8. Low T, Liu Y, Putnam F. Structure, function, and evolutionary relationships of Fc domains of human immunoglobulins A, G, M, and E. Science. 1976 Jan 30;191(4225):390–2.

9. Ben Mkaddem S, Benhamou M, Monteiro RC. Understanding Fc Receptor Involvement in Inflammatory Diseases: From Mechanisms to New Therapeutic Tools. Front Immunol. 2019 Apr 12;10.

10. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 2012;13(5):363–73.

11. Lefranc MP. Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of Immunoinformatics. Front Immunol. 2014;5.

12. Brusco A, Cinque F, Saviozzi S, Boccazzi C, DeMarchi M, Carbonara AO. The G4 gene is duplicated in 44% of human immunoglobulin heavy chain constant region haplotypes. Hum Genet. 1997 Jul;100(1):84–9.

13. Bashirova AA, Zheng W, Akdag M, Augusto DG, Vince N, Dong KL, et al. Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) genes. Genes Immun. 2021 Dec;22(7):327–34.

14. Calonga-Solís V, Malheiros D, Beltrame MH, Vargas L de B, Dourado RM, Issler HC, et al. Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection. Front Immunol. 2019 Jun;10:1161.

15. Richardson SI, Lambson BE, Crowley AR, Bashirova A, Scheepers C, Garrett N, et al. IgG3 enhances neutralization potency and Fc effector function of an HIV V2-specific

broadly neutralizing antibody. PLoS Pathog. 2019 Dec;15(12):e1008064.

16. Jefferis R, Lefranc MP. Human immunoglobulin allotypes. mAbs. 2009;1(4):332–8.

17. Lefranc MP, Lefranc G. Human Gm, Km, and Am allotypes and their molecular characterization: a remarkable demonstration of polymorphism. Methods Mol Biol Clifton NJ. 2012;882:635–80.

18. Carayannopoulos L, Hexham JM, Capra JD. Localization of the binding site for the monocyte immunoglobulin (Ig) A-Fc receptor (CD89) to the domain boundary between Calpha2 and Calpha3 in human IgA1. J Exp Med. 1996 Apr 1;183(4):1579–86.

19. Pleass RJ, Dunlop JI, Anderson CM, Woof JM. Identification of residues in the CH2/CH3 domain interface of IgA essential for interaction with the human fcalpha receptor (FcalphaR) CD89. J Biol Chem. 1999 Aug 13;274(33):23508–14.

20. Shields RL, Namenuk AK, Hong K, Meng YG, Rae J, Briggs J, et al. High resolution mapping of the binding site on human IgG1 for Fc gamma RI, Fc gamma RII, Fc gamma RIII, and FcRn and design of IgG1 variants with improved binding to the Fc gamma R. J Biol Chem. 2001 Mar 2;276(9):6591–604.

21. Yamane-Ohnuki N, Kinoshita S, Inoue-Urakubo M, Kusunoki M, Iida S, Nakano R, et al. Establishment of FUT8 knockout Chinese hamster ovary cells: an ideal host cell line for producing completely defucosylated antibodies with enhanced antibody-dependent cellular cytotoxicity. Biotechnol Bioeng. 2004 Sep 5;87(5):614–22.

22. Lazar GA, Dang W, Karki S, Vafa O, Peng JS, Hyun L, et al. Engineered antibody Fc variants with enhanced effector function. Proc Natl Acad Sci U S A. 2006 Mar 14;103(11):4005–10.

23. Stavenhagen JB, Gorlatov S, Tuaillon N, Rankin CT, Li H, Burke S, et al. Fc optimization of therapeutic antibodies enhances their ability to kill tumor cells in vitro and controls tumor expansion in vivo via low-affinity activating Fcgamma receptors. Cancer Res. 2007 Sep 15;67(18):8882–90.

24. Nordstrom JL, Gorlatov S, Zhang W, Yang Y, Huang L, Burke S, et al. Anti-tumor activity and toxicokinetics analysis of MGAH22, an anti-HER2 monoclonal antibody with enhanced Fcγ receptor binding properties. Breast Cancer Res BCR. 2011;13(6):R123.

25. Lee PD, Mukherjee S, Edeling MA, Dowd KA, Austin SK, Manhart CJ, et al. The Fc Region of an Antibody Impacts the Neutralization of West Nile Viruses in Different Maturation States. J Virol. 2013 Dec;87(24):13729–40.

26. Mimoto F, Igawa T, Kuramochi T, Katada H, Kadono S, Kamikawa T, et al. Novel asymmetrically engineered antibody Fc variant with superior FcγR binding affinity and specificity compared with afucosylated Fc variant. mAbs. 2013;5(2):229–36.

27. Ahmed AA, Keremane SR, Vielmetter J, Bjorkman PJ. Structural characterization of GASDALIE Fc bound to the activating Fc receptor FcγRIIIa. J Struct Biol. 2016 Apr;194(1):78–89.

28. Krah S, Sellmann C, Rhiel L, Schröter C, Dickgiesser S, Beck J, et al. Engineering bispecific antibodies with defined chain pairing. New Biotechnol. 2017 Oct 25;39(Pt B):167–73.

29. Krah S, Kolmar H, Becker S, Zielonka S. Engineering IgG-Like Bispecific Antibodies—An Overview. Antibodies. 2018 Sep;7(3):28.

30. Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies. 2019 Dec 3;8(4):55.

31. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016 Oct;4:e2584.

32. Jana U, Rodriguez OL, Lees W, Engelbrecht E, Vanwinkle Z, Peres A, et al. The human immunoglobulin heavy chain constant gene locus is enriched for large complex structural variants and coding polymorphisms that vary in frequency among human populations. bioRxiv: the preprint server for biology [Internet]. 2025 Dec;2025.02.12.634878.

33. Hexham JM, White KD, Carayannopoulos LN, Mandecki W, Brisette R, Yang YS, et al. A Human Immunoglobulin (Ig)A Cα3 Domain Motif Directs Polymeric Ig Receptor–mediated Secretion. J Exp Med. 1999 Feb 15;189(4):747–52.

34. White KD, Capra JD. Targeting Mucosal Sites by Polymeric Immunoglobulin Receptor-directed Peptides. J Exp Med. 2002 Aug 19;196(4):551–5.

35. Damelang T, de Taeye SW, Rentenaar R, Roya-Kouchaki K, de Boer E, Derksen NIL, et al. The Influence of Human IgG Subclass and Allotype on Complement Activation. J Immunol Author Choice. 2023 Dec 1;211(11):1725–35.

36. de Taeye SW, Bentlage AEH, Mebius MM, Meesters JI, Lissenberg-Thunnissen S, Falck D, et al. FcγR Binding and ADCC Activity of Human IgG Allotypes. Front Immunol. 2020 May 6;11:740.

37. Ford EE, Tieri D, Rodriguez OL, Francoeur NJ, Soto J, Kos JT, et al. FLAIRR-Seq: A Method for Single-Molecule Resolution of Near Full-Length Antibody H Chain Repertoires. The Journal of Immunology. 2023 Apr 7;210(10):1607–19.

38. Rispens T, Davies AM, Ooijevaar-de Heer P, Absalah S, Bende O, Sutton BJ, et al. Dynamics of Inter-heavy Chain Interactions in Human Immunoglobulin G (IgG) Subclasses Studied by Kinetic Fab Arm Exchange. J Biol Chem. 2014 Feb 28;289(9):6098–109.

39. Lewis MJ, Pleass RJ, Batten MR, Atkin JD, Woof JM. Structural requirements for the interaction of human IgA with the human polymeric Ig receptor. J Immunol Baltim Md 1950. 2005 Nov 15;175(10):6694–701.

40. Turula H, Wobus CE. The Role of the Polymeric Immunoglobulin Receptor and Secretory Immunoglobulins during Mucosal Infection and Immunity. Viruses. 2018 May 3;10(5):237.

41. Vidarsson G, Dekkers G, Rispens T. IgG Subclasses and Allotypes: From Structure to Effector Functions. Front Immunol. 2014 Oct 20;5:520.

42. Taylor RA, Xiao S, Carias AM, McRaven MD, Thakkar DN, Araínga M, et al. PET/CT targeted tissue sampling reveals virus specific dIgA can alter the distribution and localization of HIV after rectal exposure. PLoS Pathog. 2021 Jun;17(6):e1009632.

43. Calmettes G, Weiss JN. A Quantitative Method to Track Protein Translocation between Intracellular Compartments in Real-Time in Live Cells Using Weighted Local Variance Image Analysis. Degtyar VE, editor. PLoS ONE. 2013 Dec 20;8(12):e81988.

Chapter 3:

Meta-analysis of IGLV and IGKV repertoires from three distinct populations

Abstract

The light chain of an antibody contributes to antigen binding capacity and is encoded by two distinct genetic loci, the immunoglobulin kappa (IGK) locus and immunoglobulin lambda (IGL) locus. Like the heavy chain, the variable genes (IGKV and IGLV) play a major role in antigen binding and demonstrate extensive polymorphic diversity. Prior studies on a small sample of participants from Australia and Papua New Guinea demonstrated that even genetically divergent populations show strong overlap in gene usage from a set of limited genes that dominate the expressed light chain repertoire. This has been postulated to result from specific regulatory mechanisms to limit the production of autoreactive responses. In this study, we examined the expressed IGK and IGL repertoires of six South Africa individuals and contrasted these with the expressed repertoires of 100 Norwegian participants and ten participants from the UK, obtained from public databases. Novel allele inference was performed uncovering 41 novel alleles, 38 of which were from IGKV genes. Next, we examined gene usage biases across these three different populations. There were eight genes that comprised the majority of the IGKV repertoire in Norwegian and UK samples: IGKV4-1, IGKV3-20, IGKV3-15, IGKV3-11, IGKV1-5 IGKV1-39/IGKV1D-39, IGKV2-28/IGKV2D-28, IGKV1-33/IGKV1D-33. These biases matched what was observed in the South African repertoires, with the exception of IGKV4-1, which was rarely expressed in South African samples. For the IGLV repertoires, there were two genes that dominated the repertoires across all three populations: IGLV3-1, IGLV2-14; although the strength of these biases were greater in the Norwegian population. This overlap in IGLV and IGKV gene usage bias is consistent with previous reports. Thus, biases exist despite significant levels of diversity between the European and South African populations, reinforcing the hypothesis that the light chain repertoires are constrained by regulatory mechanisms. It is important to understand these dynamics to inform how these biases may impact pathogenesis and vaccination efforts, such as germline targeting vaccine strategies which need to be effective across genetically diverse populations.

Introduction

Antibodies target pathogens through their antigen binding sites which have affinity for specific surface molecules. The antigen binding site comprises the variable regions of both the heavy and light chains of an antibody, which are encoded by highly polymorphic v-genes (1). We have previously discussed (in Chapter 1) the heavy chain variable (IGHV) genes, here we examine the light chain variable genes in the same population. The light chain can arise from one of two genetic loci in the human genome, the immunoglobulin kappa (IGK) locus on chromosome 2 or the immunoglobulin lambda (IGL) locus on chromosome 22 (2).

The genetic loci for the light chain are different in their composition from the IGH locus. The light chain entirely lacks a diversity (D) gene and has a significantly shorter constant gene, which does not define the isotype of the antibody like the IGH constant genes (3). There are 33 functional IGLV genes that are subdivided into eleven phylogenetic subgroups (4,5). There are 38 functional IGKV genes that are separated into six phylogenetic subgroups and are further divided into two largely identical clusters, proximal and distal (4). Genes from the distal IGKV cluster are designated separately with a 'D' in the gene name (i.e. IGKV1D-16) and exist in the opposite orientation to genes of the proximal IGKV gene cluster (3,6). These separate IGKV gene clusters arose through a large-scale segmental duplication. This structural arrangement is consistent across almost all human individuals, in contrast to the more recent duplication/insertion/deletion events in the IGH locus that vary largely by population (2,7).

The IGK and IGL loci also demonstrate significant polymorphic and copy number diversity, similar to what is observed within the IGH locus. However, there is less genetic diversity within both light chain loci when compared to the heavy chain locus (8,9). Far fewer alleles and structural variants have been described for IGKV and IGLV genes. This may be in part due to the trend across antibody repertoire studies to focus on the heavy chain, as evidenced by IGH repertoires taking up the large bulk of publicly available data (10,11).

Recent locus hybridization capture studies have uncovered novel allelic and structural variants in the IGK and IGL loci (12,13). In the IGK locus, 67 novel alleles were discovered across a dataset of diverse populations (12). The populations demonstrated allele frequencies that could be differentiated along demographic lines, similar to that which has been observed in the IGH locus where African populations demonstrate particular patterns of genetic variability in contrast to other populations (14). Two novel structural variants were also uncovered in the IGK loci sequenced in this study, which are large scale transversion and inversion events that

occur between the proximal and distal IGKV loci. These events were observed at higher frequencies in African populations than in other populations. Hybridisation capture was also performed on the IGL locus (13). There were 36 novel IGLV alleles uncovered in this study. However, there was only a single structural variant uncovered which was not in the variable genes but amongst the constant genes, being a duplication of IGLC3. Another recent large scale study on the expressed repertoires of 100 Norwegian participants has also uncovered novel genetic variation, with 48 novel IGLV and IGKV alleles uncovered 11 of which were validated by Sanger sequencing (15). Furthermore, the study uncovered novel variation in the leader sequence of light chain variable genes, similar to what was observed and discussed in Chapter 1 of this thesis for the heavy chain locus.

Previous studies have shown that the light chain variable genes show strong usage biases, independent of selective pressure from antigens (15–18). In general, IGKV genes are expressed more than IGLV genes, at a ratio of 2:1 (19). Furthermore, these biases in gene usage have been examined across two different populations, Australian and Papua New Guinean, where the same biases were present across both populations (17,18). For the IGKV genes, there is enrichment of seven genes: IGKV3-20, IGKV3-15, IGKV3-11, IGKV1-5, IGKV2-30, and IGKV1-39/IGKV1D-39 (17). One gene, IGKV3-20, was present in more than 30% of the repertoire for certain individuals. Furthermore, proximal cluster IGKV genes are expressed at much higher levels than distal cluster genes (20). In IGLV genes, more than 20% of lambda rearrangements were shared between different individuals; although it was uncertain from this study what genes contributed to these IGL responses. However, it must be noted that these studies were performed in only two participants from either group and with sequencing technology that provided much more limited depth than what is available currently. Therefore, it remains to be seen if studies of more participants and more populations will reveal the same biases.

There is a fundamental regulatory reason for the gene usage biases in the light chain repertoires (2,21). During B-cell development, nascent light chain rearrangements cycle through combinations of IGKV and IGKJ genes after establishment of a productive and non-autoreactive heavy chain. Combinations that have high affinity for self-antigens are discarded due to their susceptibility for autoimmune reactions (22). Once the IGK locus has been exhausted, then IGL combinations are utilised (21). This is what contributes to the prevalence of IGK expression over IGL expression and restricts the usage of particular light chain genes.

Germline targeting vaccine design is a promising avenue for the field of HIV prevention (23–25). This involves creating immunogens that bind with high affinity to B-cell precursors which exhibit traits that are present in broadly neutralizing anti-HIV lineages. One of the key traits

that often defines these lineages is the germline IG genes that produce the antibodies. An early human trial of the germline immunogen designed to elicit VRC01-like antibodies showed that the alleles present in an individual for the target gene, IGHV1-2, dramatically impacted production of the target response (26). The individual in the study that lacked either IGHV1-2*02 or IGHV1-2*04 produced no VRC01-like precursor. This indicates that the composition of the innate germline of an individual is a crucial factor in the efficacy of a germline targeting vaccine. Furthermore, it is also likely that the ability of an individual to mount the desired response to a particular germline immunogen may be affected by the expression of those germline genes (27).

The aim of this study is to analyse IGK and IGL repertoires from South African individuals for novel alleles, and to understand the expression usage bias of IGK and IGL genes. Furthermore, we compared these data to that from two other populations from the United Kingdom and Norway, contrasting the usage of genes across IGKV and IGLV repertoires. Given the gene usage biases in IGLV and IGKV genes that were observed, we also assessed the likelihood of beneficial precursors for anti-HIV lineages based on their light chain germline usage.

Methods

Participant samples and ethics

The samples used for this study are from the same transcriptomic datasets utilised in Chapter 1, detailed under Samples and Ethics. The transcriptomic datasets were produced using a 5'MTPX (MulTiPleX) approach with multiplex primers that enrich IGLV and IGKV through targeting their leader sequences (28). This study also utilises naive IGK and IGL transcriptomes from two other studies (15,29). The first is a dataset of 100 transcriptomes from a sample of Norwegian participants, termed the NOR dataset in this study, produced by bulk 5'RACE (Rapid Amplification of cDNA Ends) sequencing of naïve B-cells (15). The second is a dataset of ten individuals sampled randomly from the UK blood bank service - NHSBT, termed the UK dataset in this study, produced through 10X Genomics single-cell sequencing of naïve b-cells (29). These datasets were publicly available from SRA (UK, BioProject:PRJNA1030331) and VDJBase (NOR, VDJBase Study Project:P1).

Preprocessing of raw sequences

The raw sequencing data from the CAPRISA samples was processed and analysed using the same method outlined in Chapter 1, to infer the germline alleles present in each repertoire, inclusive of inferring novel alleles. In this study, however, instead of using IMGT reference data as a reference database for inference, we utilised a custom database from https://bitbucket.org/yaarilab/processpipeline/src/master/igblast_reference/, from Mikocziova et. al (15). The reason for this is that it accounts for the inability of inference pipelines to discriminate between identical V-regions from the proximal and distal gene clusters by assigning sequences to a unified gene name, for example IGKV1-39/IGKV1D-39 are denoted as IGKV1E-39. This avoids a skewed representation of proximal/distal gene usage or assignment. The antibody sequences from the NOR dataset were analysed in the same way that we have performed our analysis, so the processed output from the supplementary data of the paper was utilised for this study. The UK dataset was only accessible as raw reads. The raw data was obtained from SRA and preprocessed using MiXCR (30,31), utilising the same custom reference database.

Analysis of IGLV and IGKV sequences

The processed tabular files from IgDiscover and MiXCr were processed through a series of custom Julia scripts (https://github.com/athenamarsden-bio/LC_PhD_Scripts). Briefly, the scripts calculated gene usage for each light chain variable gene by determining the percentage of unique VJ sequences associated with each gene as a proportion of total unique VJ sequences. The means for gene usage in each gene was calculated for each population in these scripts as well. The tabular output from these calculations was visualised in GraphPad Prism 8. The scripts also compiled all unique allele sequences inferred into population specific files for downstream processing. The sequences were annotated by population (SA, NOR, UK) and aligned using Clustal Omega (32). IGKV and IGLV highlighter plots were created using LANL alignment plots, the tables were produced using Google Sheets. (https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html)

Results

IGKV and IGLV allelic diversity

The IGKV and IGLV expressed repertoires from six South African individuals were analysed as part of this study, each repertoire containing an average of ˜500 000 unique VDJ sequences. The repertoires came from naive B-cells, thus representing baseline gene usage biases and allelic expression. First, allele inference was performed on these sequences with a custom database. There were a total of 93 unique IGKV alleles uncovered across 35/38 functional genes. Of these alleles, 38 were novel (41%, Figure 1A and C) occurring across 22 of the 35 IGKV genes. These were observed in both proximal and distal genes (denoted with a D), however for some genes, such as IGKV1E-39, it was impossible to determine if the alleles were from distal or proximal genes. There were also alleles that matched non-IMGT alleles from a recent genomic study (n=7, 8%) (12).

There were fewer IGLV genes recovered with only 21 of the total number of 33 functional genes detected (Figure 1B and D). This was likely due to IGLV genes being less expressed in general than IGKV genes. There were 36 unique IGLV alleles observed in the dataset, three of which were determined to be novel (8%). The remaining six (17%) alleles matched non-IMGT alleles from a recent genomic study (13).

When we compared the proportions of novel alleles between the variable genes of the IG loci (IGK, IGL and IGH; Figure 1C, D and E), there was a greater proportion of novel alleles in IGK (41%) when compared to IGL (8%) and IGH (21%). The sequencing performed on naive IGKV and IGLV repertoires was performed on the same individuals from Chapter 1 with naive IGHV transcriptomes indicating that novel diversity is prevalent across all IG loci in South African individuals. Sequences from the previously mentioned locus capture studies were also identified (denoted as ‘other’ in Figure 1)(12,13). These sequences accounted for 8% of IGKV alleles and 17% of IGLV alleles.

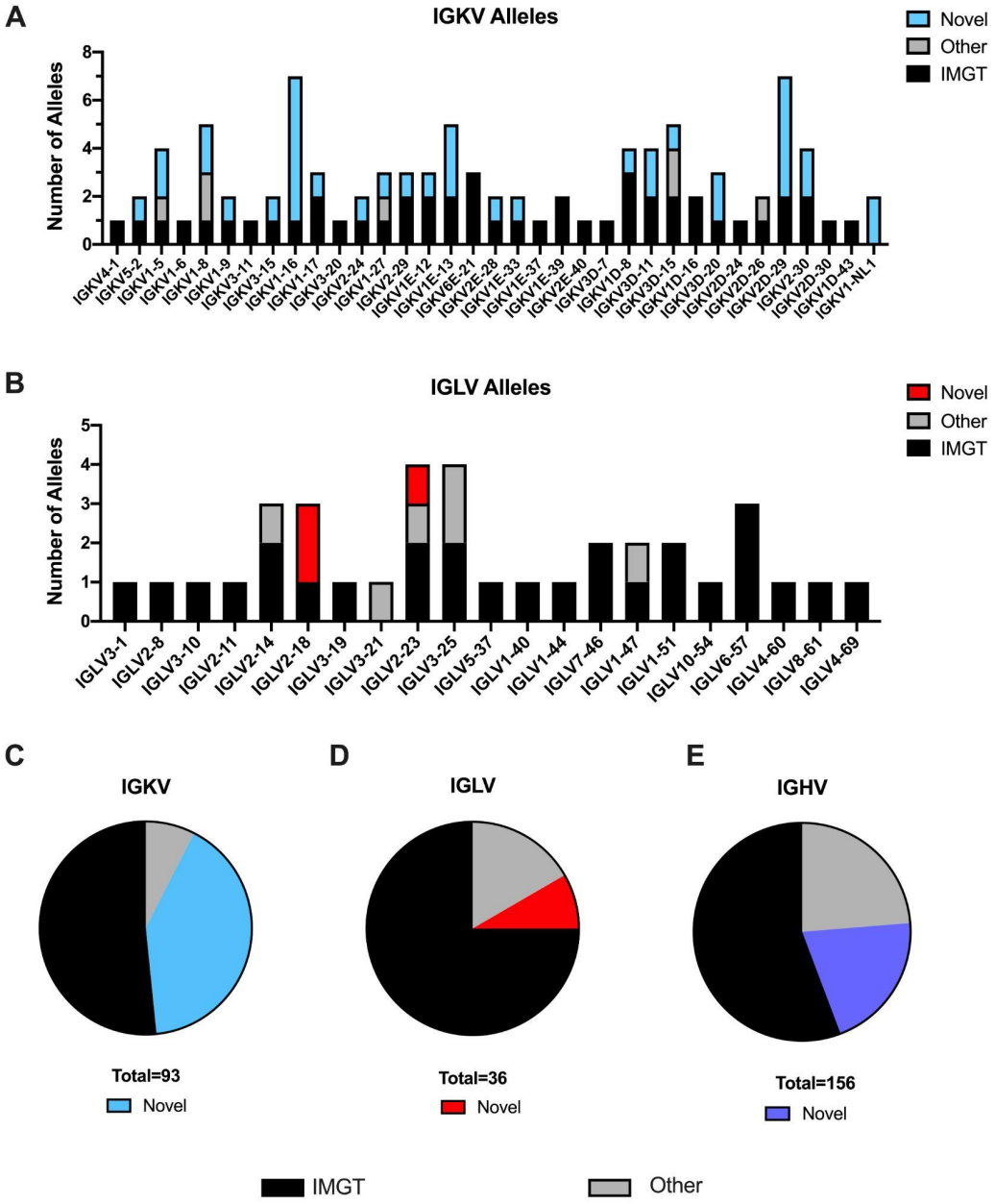


Figure 1: Distribution of IGKV and IGLV alleles. A: The number of IGKV alleles per gene that were novel (blue), matched IMGT reference data (black), or matched alleles from other studies (grey). B: The number of IGLV alleles per gene that were novel (red), matched IMGT reference data (black) and matched alleles from other studies (grey). C: The proportion of IGKV alleles that were novel (blue) in contrast to reference IMGT alleles (black), or matched alleles from other studies (grey). D: The proportion of IGLV alleles that were novel (red) matched IMGT reference data (black) and matched alleles from other studies (grey). E: From Chapter 1, the IGHV alleles that were designated as novel (purple), IMGT (black) or matched alleles reported in other databases (grey) from transcriptome sequencing.

IGKV and IGLV gene usage biases between different populations

Next, we compared the gene usage in naive light chain repertoires from these six South African repertoires with a Norwegian population (n=100, ˜30 000 unique VJs per participant) and a population based in the United Kingdom (n=10, ˜50 000 cells analysed per participant). This was done to examine whether gene usage biases are similar across different populations, as has been previously observed, or if there are certain population specific gene biases (17,18). We utilised the Norwegian dataset as the baseline for comparison, identifying the most used genes in the datasets and contrasting them with usage in the UK and SA populations. The Norwegian and UK datasets were chosen as they were the only comprehensive paired IGKV and IGLV naive repertoires publicly available.

The IGKV repertoires demonstrated strong overlap in gene usage biases between the three populations (Figure 2). When we contrast these gene usage biases with the European datasets, we observe that there is similarity between the IGKV genes that comprise the majority of the expressed IGKV repertoire (Figure 2B and C). There is greater variability in the Norwegian dataset (Figure 2B), which is the result of the sample size being significantly larger (n=100). The most prominent genes (median above 5%) in the NOR dataset were: IGKV4-1, IGKV3-20, IGKV3-15, IGKV3-11, IGKV1-5 IGKV1E-39, IGKV2E-28, IGKV1E-33. These same genes were prominent in the UK population at roughly the same prevalence (difference between the means only exceeded 4% for IGKV3-11, Figure 2F). While the same expression bias was observed for many of the genes in the SA dataset, there was a notable difference in the expression of IGKV4-1 between the South African population and the European populations. The expression of IGKV4-1 was very low (<1%) in South African populations but is much more prevalent in European populations (5%-20%). Across all data sets, distal (D) genes were lowly expressed.

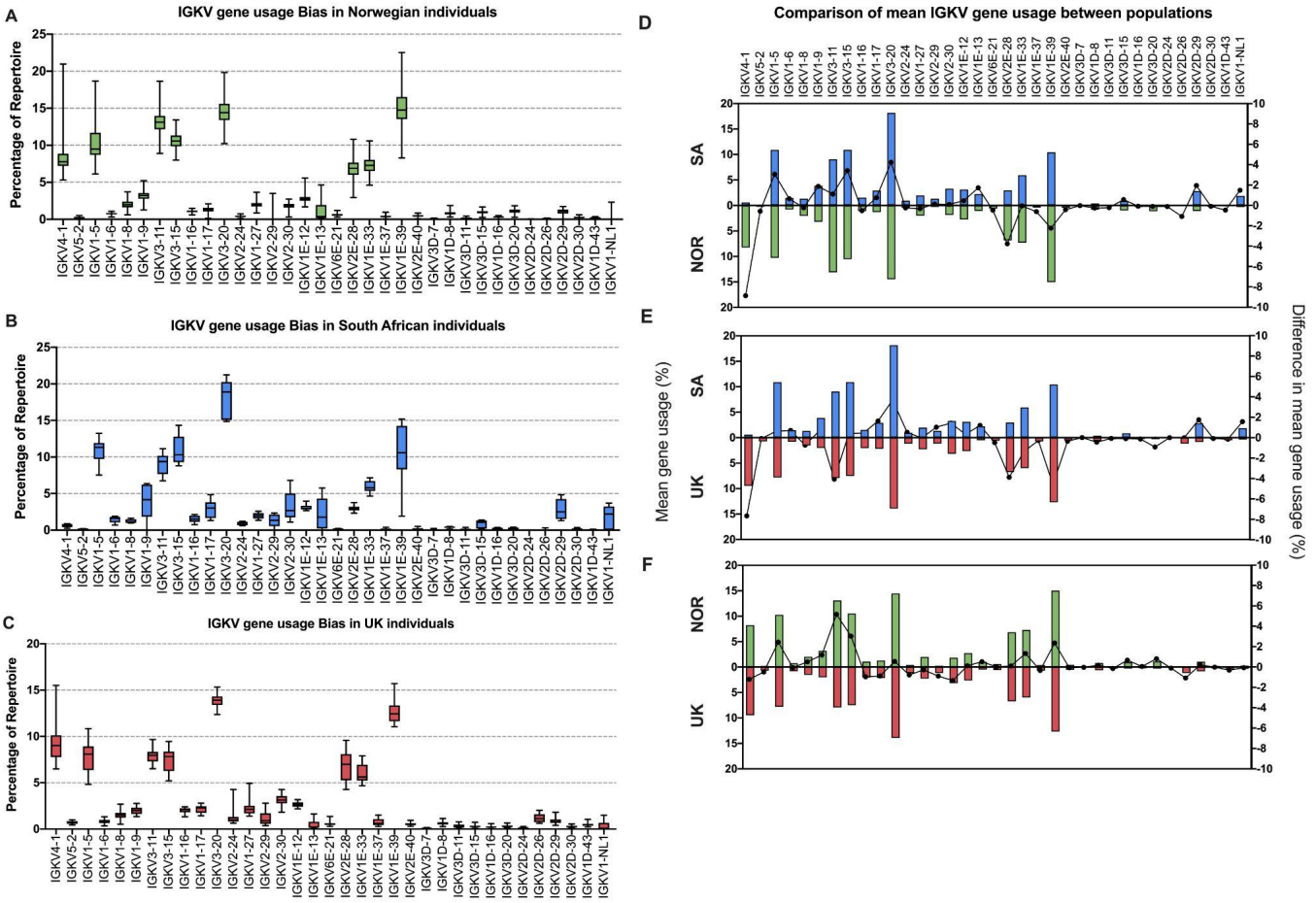


Figure 2: IGKV gene usage biases. A-C: The distribution of gene usage biases across IGKV genes in NOR (A, green), SA (B, blue) and UK (C, red) populations. The percentage of the repertoire that each gene occupies within individuals of the population is represented. D-F: The differences in mean gene usage between the three populations (black line, measured on the right).

We also assessed the genetic similarities between the three populations in this study for the eight most prominent genes identified in the NOR dataset (Figure 3). For all genes, there were more alleles present in the NOR dataset, likely due to the large sample size. Most of the differences between different alleles were single SNPs, with the exception of IGKV1-5 and IGKV1E-39 where there are more significant differences between alleles. The SA and UK datasets also expressed many of the same alleles. In IGKV4-1, the same *01 allele was present in all datasets and the only allele described for SA and UK and even NOR participants, despite the greater numbers in this cohort. This suggests the marked difference in expression for IGKV4-1 is influenced by other, possibly intergenic, changes. There are four novel alleles from the SA population which contribute to these dominant gene biases (two alleles for IGKV1-5; one for IGKV3-15, IGKV1E-33 and IGKV2E-28).

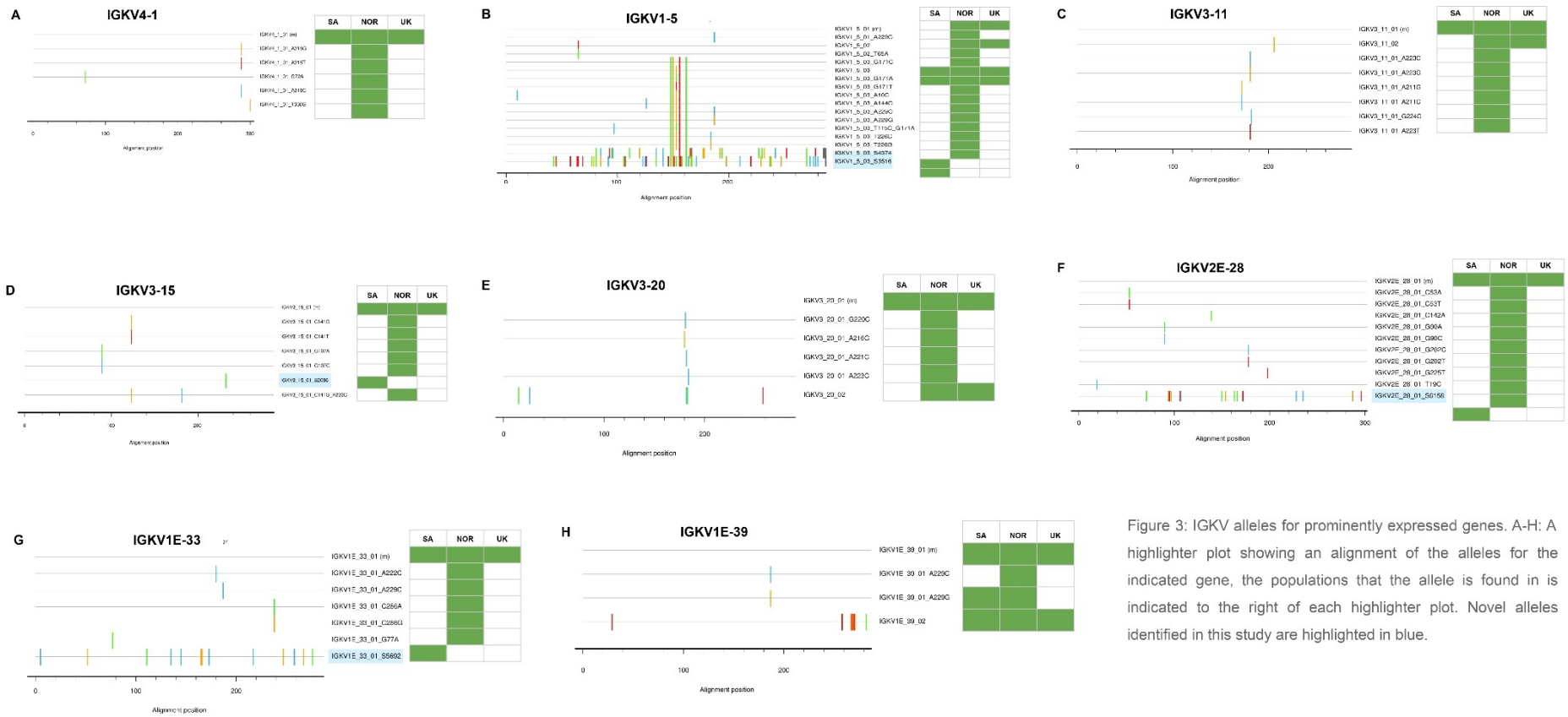


Figure 3: IGKV alleles for prominently expressed genes. A-H: A highlighter plot showing an alignment of the alleles for the indicated gene, the populations that the allele is found in is indicated to the right of each highlighter plot. Novel alleles identified in this study are highlighted in blue.

The IGLV genes also demonstrate patterns of gene usage bias between African and European populations, with these biases often being starker than what was detected in the IGKV repertoire (Figure 4). The most prominent eight IGLV genes (median above 10%) in the NOR population were: IGLV3-1, IGLV2-14, IGLV3-19, IGLV2-23, IGLV3-25, IGLV1-40 and IGLV1-44. There is one gene, IGLV3-1, that occupies between 10-65% of the IGLV repertoire in NOR individuals. This is also amongst the most prevalent IGLV genes in the SA dataset (25-28%) and UK dataset (9%-18%). The next most prominent gene, IGLV2-14, is also prevalent across the three datasets: SA - (9%-15%), NOR - (18%-68%), UK - (7%-14%). These two genes account for a significant portion of the repertoire in these datasets, with other genes contributing less to the repertoire. The biases in these two genes appear to be stronger in the NOR repertoires, although this seems to be common across all IGLV genes, with many genes having a mean bias difference greater than ten for both the UK and SA repertoires (Figure 4D and F). The third most prominent gene in the SA dataset is IGLV3-25 (10%-15%), this is distinct however from the European repertoires where IGLV3-25 is less prominent than other genes such as IGLV1-40. Overall, there are two genes that dominate the IGLV repertoire (IGLV3-1, IGLV2-14) and this is consistent between the different populations, while other genes do show population differences.

Genetically, IGLV genes show less allelic variation than in IGKV, with the exception of IGLV2-14, where NOR alleles show significant nucleotide variation (Figure 5). Interestingly, IGLV3-1 which dominates the IGLV repertoire in all populations only has five alleles present in the data set, with very few SNPs that differentiate them. As with the IGKV alleles, there is strong overlap between the alleles present between the SA and UK datasets. There are two genes with significant expression that expressed novel alleles in the SA population, IGLV2-23 and IGLV3-25.

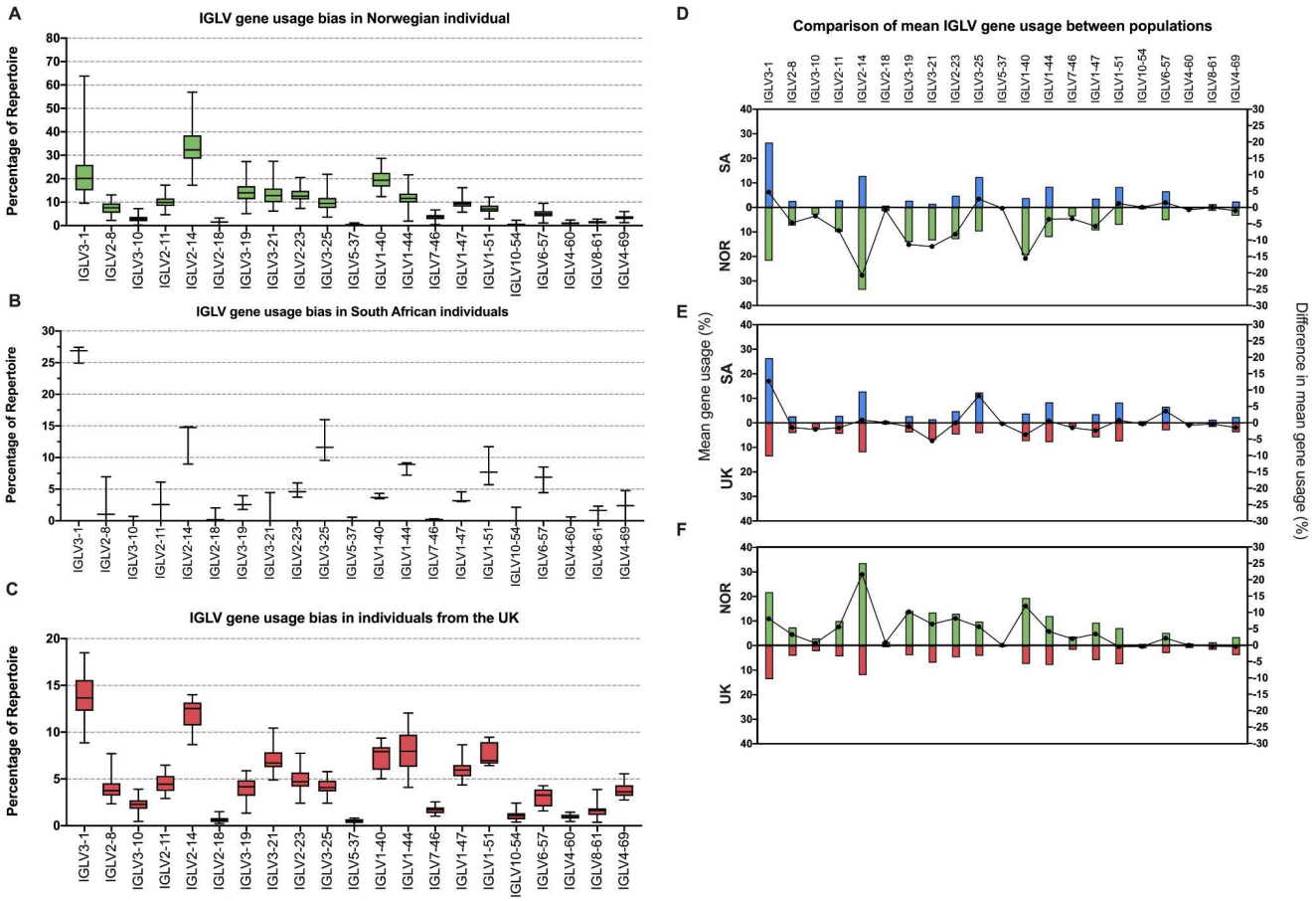


Figure 4: IGLV gene usage biases. A-C: The distribution of gene usage biases across IGLV genes in NOR (A, green), SA (B, blue) and UK (C, red) populations. The percentage of the repertoire that each gene occupies within individuals of the population is represented. D-F: The differences in mean gene usage between the three populations (black line, measured on the right).

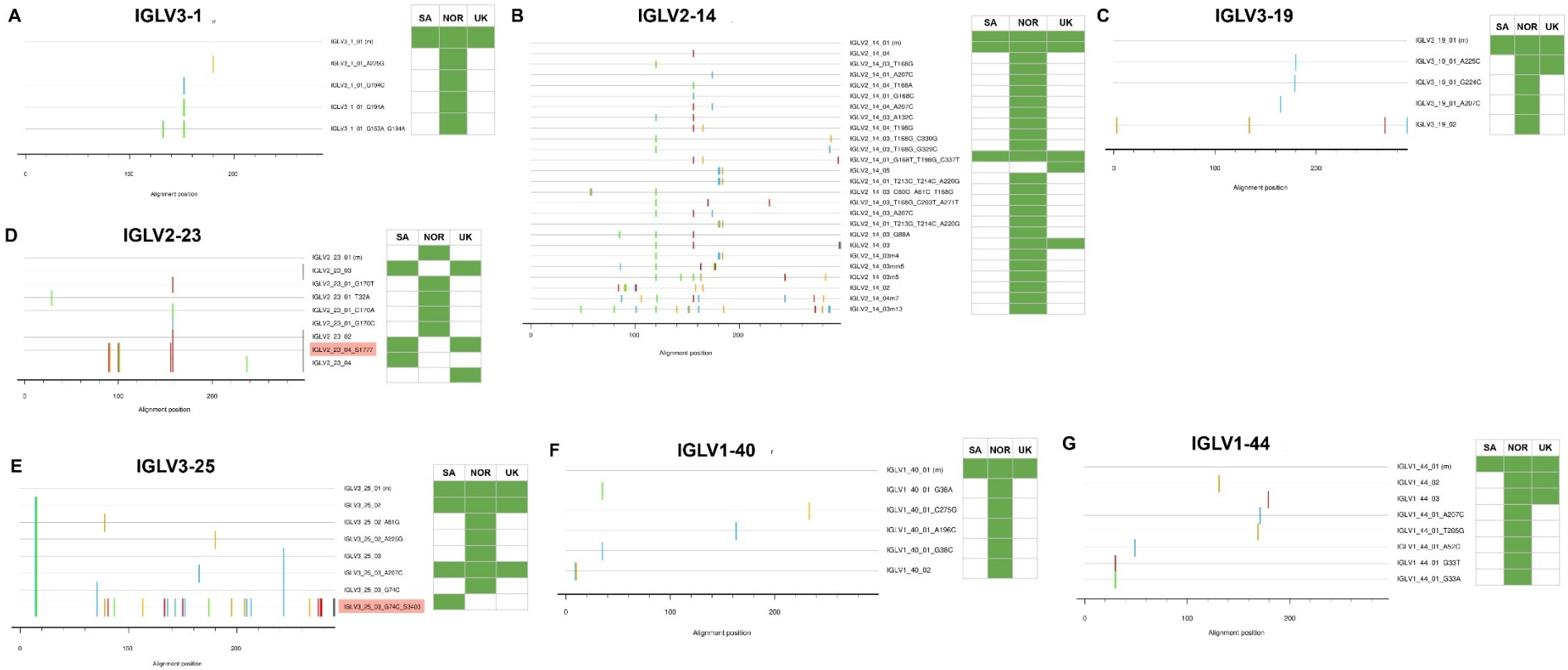


Figure 5: IGLV alleles for prominently expressed genes. A-H: A highlighter plot showing an alignment of the alleles for the indicated gene, the populations that the allele is found in is indicated to the right of each highlighter plot. Novel alleles are highlighted in red.

Expression of genes associated with public anti-HIV antibodies for germline targeting immunogen design

For germline targeting strategies to have the best chance of eliciting the desired precursors, it would be beneficial to target lineages with genes that show high levels of expression across a variety of populations. For example, there are five bNAb lineages that use IGKV3-20 (VRC01, 1-18, PCDN76-33A, CH01 and PCT64-35Md; Table 1), of which the VRC01-like class produces especially broad and potent antibodies across different individuals and has been demonstrated to be an effective target for germline targeting design (26,33). Across all three populations that were analysed in this study, IGKV3-20 is a prominently expressed gene, occurring at mean prevalence between 14% and 18%. Another IGKV gene, IGKV1-33, part of three bNAb lineages (12A12, 3BNC117, VRC-CH31; Table 1), demonstrates mean prevalence around 6% across the three populations and may prove to be effective as a target of germline vaccine strategies. In contrast, PCIN63.71I is derived from IGKV1-6 which occurs at low prevalence in all populations (1.5%-0.8%) and is therefore unlikely to be an effective candidate for immunogen design (Table 1).

The lambda genes might be less useful targets across different populations and individuals given that we and others have demonstrated that gene usage biases are strongly restricted to two IGLV genes (17,18). The most prominent IGLV3-1 gene (14%-25% mean prevalence across study populations) is only utilised by one bNAb lineage (CH103). The PGT121 bNAb lineage, which is derived from IGLV3-21, is rarely expressed in the SA dataset but shows reasonable expression in the UK and NOR datasets. These examples demonstrate that more knowledge is needed about the germline expression dynamics across different populations and how they may inform germline targeting strategies. The ability of a germline targeting immunogen to elicit bNAb responses has, in prior studies, focussed predominantly on the heavy chain gene usage. While it is likely that IGHV germline variants, that we discuss in Chapter 1 have implications for germline targeting vaccine strategies, we focus in this thesis on expanding the analysis to include variation in the light chain variable genes as this is under-studied.

Table 1: HIV bNAb lineages prioritised for germline targeting immunogens

Adapted from Haynes et. al (25). Multiple light chain genes are also coloured. Prevalence in the repertoire is shown for the light chain gene used. The legend for colouring at the bottom of the table.

HIV-1 epitope	bnAb	Breadth (%)	Potency (μg ml–1)	VH gene	VL gene	Mean Prevalence in SA repertoire	Mean Prevalence in NOR repertoire	Mean Prevalence in UK Repertoire	Germline-targeting priming immunogen
CD4 Binding Site	12A12	93	0.221	VH1–2	VK1–33	5.92	5.97	7.30	eOD-GT8
	VRC01	91	0.377	VH1–2	VK3–20	18.16	13.92	14.47	426c.TM4ΔV1-3 , BG505 SOSIP 4.1 GT1, eOD-GT8
	3BNC117	89	0.116	VH1–2	VK1–33	5.92	5.97	7.30	eOD-GT8
	VRC-CH31	84	0.321	VH1–2	VK1–33	5.92	5.97	7.30	426c.TM4ΔV1-3, eOD-GT8
	PCIN63.71I	84	0.46	VH1–2	VK1–6	1.462	0.812	0.753	
	IOMA	49	2.33	VH1–2	VL2–23	6.95	12.96	4.13	426c.TM4ΔV1-3, eOD-GT8
	1-18	97	0.048	VH1–46	VK3–20	18.16	13.92	14.47	
	CH235.12	89	0.7	VH1–46	VK3–15	10.91	7.50	10.52	
	CH103	67	2.28	VH4–59	VL3–1	24.92	21.74	13.66	CH505 T/F33
V3 Glycan	PGT121	66	0.072	VH4–59	VL3–21	>0.01	13.43	7.00	MD39-11MUTb, RC1-4Fill
	BG18	61	0.032	VH4–4	VL3–25	15.99	9.79	4.16	MD39-11MUTb, N332 GT5
	BF520.1	53	7.31	VH1–2	VK3–15	10.91	7.50	10.52	
	PGDM12	54	0.14	VH3–11	VK2–24	0.93	1.18	0.40	
	DH270.6	51	0.21	VH1–2	VL2–23	6.95	12.96	4.13	CH848 SOSIP 10.17 DT
	PCDN76-33A	46	0.5	VH4–34	VK3–20	18.16	13.92	14.47	
V2 apex	PG9	87	0.154	VH3–33	VL2–14	>0.01	33.62	0.65	BG505 SOSIP 4.1 GT1, BG505 SOSIP 4.1 GT1.1 , ZM233 , KER2018, BB201.B42
	PGDM1400	83	0.02	VH1–8	VK2–28	2.98	6.75	6.86	
	VRC26.25	70	0.004	VH3–30	VL1–51	11.72	7.15	7.59	
	CH01	54	1.38	VH3–30	VK3–20	18.16	13.92	14.47	BG505 SOSIP 4.1 GT1, BG505 SOSIP 4.1 GT1.1, ZM233, CM244 Q23.17 , WITO132, T250
	PCT64-35Md	35	0.41	VH3–15	VK3–20	18.16	13.92	14.47	
Breadth	**0-25%**	**25-50%**	**50-75%**	**75-100%**					
Potency	**0-0.5**	**0.5-1.0**	**1.0-1.5**	**1.5-2.0**	**2-2.5**	**>2.5**			
Mean	**0-5%**	**5-10%**	**10-15%**	**15-20%**	**20-25%**	**25-30%**	**30-35%**		

Discussion

In this study we uncovered novel IGKV and IGLV alleles, with a majority of these found in IGKV genes. While it has been suggested that IGKV and IGLV genetic diversity is almost fully categorised, this study and other recent studies indicate this not to be the case (2,9,12,15). As more populations across the globe are sequenced with modern techniques, both genomically and transcriptomically, it's very likely that more novel alleles will be uncovered. Specifically, the small South African population used in this study cannot represent broader African diversity and thus other populations need to be examined to build a more comprehensive understanding of African diversity in the light chain loci.

In total we identified 41 novel alleles, 38 of which were from the IGKV genes. Of these novel alleles, three IGKV and four IGLV sequences matched novel alleles uncovered in prior locus capture sequencing studies (12,13). This represents a significant contribution to global immunogenetic databases as the variation in the light chain genes is relatively understudied, compared to the IGHV genes. The study was performed on six people and it is likely that these variants are common in a larger population. Due to this limitation in size, it is difficult to ascertain the frequency at which these and other variants are present in Southern African populations. Thus, it is likely that more rare variants are present in the broader population and these may have an impact on usage of IGKV and IGLV genes.

While we were able to recover sequences from the majority (35 out of 38) of IGKV genes, there were 12 missing IGLV genes in the repertoires of the study population. This may be due to the fact that IGKV genes are much more readily expressed as light chains and due to the restriction of IGLV genes where only two genes dominate the repertoire, it may require more depth in future studies to recapitulate IGLV genes that expressed at significantly low levels.

Previous studies have shown that light chain repertoires demonstrate particular patterns of gene usage bias (2,17,18). In this meta-analysis, we provide updated evidence with modern techniques and more diverse datasets that support these findings. The IGKV repertoire demonstrated strong overlap in the bias of IGKV3-20, IGKV3-15, IGKV3-11, IGKV1-5, and IGKV1-39/IGKV1D-39 between European and African populations. Furthermore, we confirm that distal IGKV genes are significantly less expressed than proximal IGKV genes which is consistent with existing literature (20). In the IGLV repertoire, two genes dominated the repertoire, IGLV3-1 and IGLV2-14, across all datasets, concordant with previous studies. These strong patterns of gene usage bias existed despite genetic differences between the

alleles that were present between African and European participants. However, for some genes such as IGKV4-1, despite genetic similarity there were substantial differences in the expression between African and European samples. It is possible that these differences result from genetic variation that exists in other parts of the locus as it has been demonstrated that intergenic variation can affect the expression of IG genes (34,35). Overall, IGLV genes showed much stronger differences in expression between the Norwegian population and the South African and UK populations. This may be partially explained by the large difference in sample size but also may be indicative of population specific divergence that has yet to be categorised, such as an undiscovered structural variant that affects the IGL locus. Furthermore, there are differences in the unique VJ counts between the three datasets, which were also produced through distinctly different means. While this may impact the comparison between the datasets, the strong concordance between gene usage identified and how this analysis reaffirms pre-existing data suggests that this is likely not the case.

In this study, as opposed to other chapters in this thesis, only transcriptomic methods were utilised. While germline methods are a more comprehensive way to examine IG genetic diversity, here we describe a significant number of novel variants demonstrating the strength of transcriptomic sequencing in immunogenetic study. Sequencing expressed repertoires also allows for the examination of gene usage biases which is a key finding in this chapter and provides information on the relationship between genetic diversity and expressed antibody repertoire, showing here that despite genetic differences between populations there is significant overlap in gene usage between these distinct populations. Furthermore, while the Norwegian used in this study had been used in the original publication to also examine allelically encoded biases, where certain alleles are expressed more readily than others, the South African dataset lacked the participants to sufficiently do a powered analysis of allelic expression. However, future datasets can be generated with more individuals to examine this as understanding more nuances of what factors influence expression is vital for connecting genetic diversity with the functional antibody response.

Understanding the dynamics of gene usage bias in light chain repertoires is important as it informs strategies such as germline targeting vaccine design. Here we have demonstrated that certain lineages such as VRC01 are likely better candidates for population agnostic vaccine strategies as the germline IGKV gene that it is derived from is highly expressed in all three populations that were examined. However, targeting of other genes may produce a less predictable response for all populations. More studies that examine the bias in naive antibody repertoires are needed to fully understand and characterise the nuances inherent in different demographics. These studies would need to include IGHV gene biases as well to create a comprehensive understanding of how population variation may impact the formation of

responses to germline targeting vaccination. We have previously demonstrated that there are a variety of potential factors that can and do impact IGHV gene expression in Chapter 1, it is likely that a combination of these factors and the constrained light chain repertoire will need to be considered together across different populations.

References

1. Lefranc MP, Lefranc G. The Immunoglobulin Factsbook. 2001. 72–239 p.

2. Collins AM, Watson CT. Immunoglobulin Light Chain Gene Rearrangements, Receptor Editing and the Development of a Self-Tolerant Antibody Repertoire. Front Immunol. 2018 Oct;9:2249.

3. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, et al. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. Genes Immun. 2015 Jan;16(1):24–34.

4. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 2009 Jan;37(Database issue):D1006-12.

5. Lefranc MP. Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of Immunoinformatics. Front Immunol. 2014;5.

6. Kawasaki K, Minoshima S, Nakato E, Shibuya K, Shintani A, Schmeits JL, et al. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. Genome Res. 1997 Mar;7(3):250–61.

7. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 2012;13(5):363–73.

8. Schwartz GW, Hershberg U. Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics. Phys Biol. 2013 Jun;10(3):035005.

9. Collins AM, Wang Y, Singh V, Yu P, Jackson KJ, Sewell WA. The reported germline repertoire of human immunoglobulin kappa chain genes is relatively complete and accurate. Immunogenetics. 2008 Nov;60(11):669–76.

10. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, et al. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev. 2018 Jul;284(1):24–41.

11. Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT, et al. VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Res. 2020 Jan 8;48(D1):D1051–6.

12. Engelbrecht E, Rodriguez OL, Shields K, Schultze S, Tieri D, Jana U, et al. Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. Genes Immun. 2024 Aug;25(4):297–306.

13. Gibson WS, Rodriguez OL, Shields K, Silver CA, Dorgham A, Emery M, et al. Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. Genes Immun. 2023 Feb;24(1):21–31.

14. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. Front Immunol. 2020 Sep 23;11:2136.

15. Mikocziova I, Peres A, Gidoni M, Greiff V, Yaari G, Sollid LM. Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes. iScience. 2021 Sep;103192.

16. Hoi KH, Ippolito GC. Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire. Genes Immun. 2013 Jun;14(4):271–6.

17. Jackson KJL, Wang Y, Gaeta BA, Pomat W, Siba P, Rimmer J, et al. Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. Immunogenetics. 2012 Jan;64(1):3–14.

18. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol Baltim Md 1950. 2010 Jun 15;184(12):6986–92.

19. Molé CM, Béne MC, Montagne PM, Seilles E, Faure GC. Light chains of immunoglobulins in human secretions. Clin Chim Acta Int J Clin Chem. 1994 Jan 31;224(2):191–7.

20. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, et al. Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. Front Immunol. 2019 Mar;10:435.

21. Langerak AW, Nadel B, De Torbal A, Wolvers-Tettero ILM, van Gastel-Mol EJ, Verhaaf B, et al. Unraveling the consecutive recombination events in the human IGK locus. J Immunol Baltim Md 1950. 2004 Sep 15;173(6):3878–88.

22. Wardemann H, Hammersen J, Nussenzweig MC. Human autoantibody silencing by immunoglobulin light chains. J Exp Med. 2004 Jul 19;200(2):191–9.

23. Sangesland M, Ronsard L, Kazer SW, Bals J, Boyoglu-Barnum S, Yousif AS, et al. Germline-Encoded Affinity for Cognate Antigen Enables Vaccine Amplification of a Human Broadly Neutralizing Response against Influenza Virus. Immunity. 2019 Oct;51(4):735-749.e8.

24. Stamatatos L, Pancera M, McGuire AT. Germline-targeting immunogens. Immunol Rev. 2017 Jan;275(1):203–16.

25. Haynes BF, Wiehe K, Borrow P, Saunders KO, Korber B, Wagh K, et al. Strategies for HIV-1 vaccines that induce broadly neutralizing antibodies. Nat Rev Immunol. 2023 Mar;23(3):142–58.

26. Leggat DJ, Cohen KW, Willis JR, Fulp WJ, deCamp AC, Kalyuzhniy O, et al. Vaccination induces HIV broadly neutralizing antibody precursors in humans. Science. 2022 Dec 2;378(6623):eadd6502.

27. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci Rep. 2016 Aug;6(1):20842.

28. Vázquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M, et al. High-Quality Library Preparation for NGS-Based Immunoglobulin Germline Gene Inference and Repertoire Expression Analysis. Frontiers in Immunology. 2019 Apr 5;10.

29. Lê Quý K, Chernigovskaya M, Stensland M, Singh S, Leem J, Revale S, et al. Benchmarking and integrating human B-cell receptor genomic and antibody proteomic profiling. Npj Syst Biol Appl. 2024 Jul 12;10(1):1–16.

30. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017 Oct 11;35(10):908–11.

31. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015 May;12(5):380–1.

32. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct 11;7:539.

33. Cohen KW, De Rosa SC, Fulp WJ, deCamp AC, Fiore-Gartland A, Mahoney CR, et al. A first-in-human germline-targeting HIV nanoparticle vaccine induced broad and publicly targeted helper T cell responses. Sci Transl Med. 2023 May 24;15(697):eadf3309.

34. Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. Nat Commun. 2023 Jul 21;14(1):4419.

35. Kenter AL, Watson CT, Spille JH. Igh Locus Polymorphism May Dictate Topological Chromatin Conformation and V Gene Usage in the Ig Repertoire. Front Immunol. 2021 May;12:682589.

Conclusion

Summary discussion of thesis findings

In this thesis, we have examined genetic diversity and its implications across several different Immunoglobulin genes, specifically focusing on those that contribute to antigen binding (IGHV, IGKV, and IGLV) and those that engage effector cells within the immune response (IGHA, IGHG). Overall, 118 alleles uncovered in this thesis were novel, marking a significant contribution to our understanding of immunogenetic diversity as well as expanding and diversifying international databases. We also described novel gene duplications in 11 IGHV genes, in which CNV is common. The novel allelic variation in the IGHA and IGHG results in amino acid changes within the Fc region of antibodies, some of which have been shown to impact the effector function of antibodies by altering Fc receptor binding or the structure of an antibody. We also demonstrated that despite significant genetic diversity between three distinct (Norwegian, South African, and United Kingdom) populations in the expressed IGKV and IGLV repertoires, there was strong overlap in the usage of a limited set of genes that dominate light chain repertoires.

While this study sequenced relatively few samples (70 for genomic sequencing and 8 for transcriptomic sequencing), the allelic variants uncovered in this study contribute to the growing inclusive and diverse databases of immunogenetic diversity as more studies are undertaken in non-European populations (1–9). Historically, much of the immunogenetic diversity recorded in public databases was from individuals of European descent (10–12). Although more studies are examining other populations, many of the largest datasets in these public databases are still European (13–15). While we examined African individuals in this thesis, the cohort comes from a specific region of South Africa, which is extremely unlikely to represent all novel diversity in African populations. It remains to be seen if immunogenetic variation follows the broader complex inter-population architecture that has been mapped out across the continent in recent years (16,17). Southern African populations cluster distinctly from other African populations due to admixture of Bantu-speaking people with older populations that were residing in the region before the Bantu migration southward, such as the San (16–18). Therefore, certain variants uncovered in the IG genes of this study population may be absent in populations outside of Southern Africa.

The majority of the genetic diversity described in this thesis is allelic, involving nucleotide variations between variants of the same gene. We were also able to detect differences in copy

number variation within the IGHV genes in the form of gene duplication, which is an instance of structural variation. Structural variation is a prevalent form of genetic diversity in the IG loci (12,19). The transcriptomic and genomic methods in this thesis can partially examine structural variation through haplotype reconstruction and estimating duplication through unique allele counts. However, these methods are unable to fully assess structural variation, as these methods cannot span the large genomic regions which encompass the structural variants. Alternative analytical tools such as read depth analysis could be used to examine CNV in the datasets presented in this thesis, however these tools are designed for WGS (whole genome sequencing) data (20). This presents a significant obstacle to implementing these approaches for our amplicon-based genomic and transcriptomic datasets. Indeed, for the transcriptomic datasets, usage biases (as detailed in Chapter 3) would prevent normalisation of the sequencing data for read depth analysis, as reads per gene are not simply influenced by CNV but by multiple other factors such as B-cell regulation and chromatin conformation. While tools exist for read depth analysis of amplicon-based genomic datasets (such as those produced in Chapter 3), they require optimisation to eliminate primer biases that may influence read depths which was not performed to ensure sequencing at sufficient depth to capture of all alleles present (21). Despite these limitations, future experiments can be designed around downstream read depth analysis as a method to examine CNV in further studies by reducing primer biases. However, new long-read sequencing techniques that can sequence the entire IG loci provide an approach that can adequately examine structural variation within the IG loci and also provide information on nucleotide variability that exists in intergenic regions (3–5). These approaches can overcome the high variability and sequence homology of the IG loci, which has proved a significant obstacle for many traditional sequencing and computational approaches. Techniques like this could yield more data about the genetic composition of the IG loci in African populations. IG loci sequencing has robust bioinformatic development and support that is publicly available to analyse the data; however, the laboratory component of this approach necessitates expert training and specialised equipment.

Beyond describing novel genetic variation, we also examined some of the functional implications that nucleotide diversity may have on the antibody response. We examined gene usage biases of the variable genes from both the heavy and light chains. It has been shown that duplication increases the expression for IGHV1-69, which was beneficial in the context of influenza vaccination, and this may hold true for other IGHV genes (22). However, it appears that duplication alone is not enough to explain how IGHV genes may have enhanced expression. Several studies have shown that IGHV gene expression is impacted by chromatin conformation (6,23,24). The looping structure of the IGH locus is highly regulated and it has

been demonstrated that mutations can affect the way that chromatin loops form. This alters the accessibility of IG genes to various enzyme complexes, enhancing or inhibiting gene expression (25,26). This has been studied in mice, through epigenetic sequencing of pro-B-cells from the bone marrow. While it is ethically inconceivable to reproduce in human studies, induced pluripotent stem cells that have been differentiated into pro-B-cells may be an approach to examine looping dynamics across diverse populations (27).

Furthermore, leader sequence mutation, shown in Chapter 1, has been utilised in commercial antibody production to enhance expression *in vitro* (28,29). It is part of the optimisation procedures in commercial production of antibodies to ensure effective expression of mAbs in bacteria or yeast platforms by altering the leader sequence. This could also explain why certain alleles or genes are expressed more readily, but this needs to be explored in vivo by more comprehensive and large-scale studies that examine the expression of allelic or leader variants across different individuals.

In contrast for the light chain loci, IGKV and IGLV genes of three distinct populations, there was a heavily constrained pattern of gene usage bias where more than 50% of the IGKV repertoire was derived from five genes and 20-50% of the IGLV repertoire was dominated by two genes. These data reinforce previous studies showing that relatively few light chain variable genes comprise the majority of the expressed IGKV and IGLV repertoire (30–32). While there were some population differences for some genes, these biases persisted across the geographically and genetically disparate populations examined in this thesis (Norwegian, South African, and from the UK). Thus, while the IGH locus has variability in expression due to a range of genetic factors, the light chain loci demonstrate consistency in gene usage. This is due to regulatory mechanisms during B-cell development which restricts light chains that are prone to self-reactivity (32,33). The necessity of these mechanisms is further supported by the altered gene usage observed in autoimmune conditions, where typically rare IGKV genes are expressed at significantly higher levels (34,35).

The allelic diversity in the constant IGHG and IGHA genes was linked to allotypic diversity, uncovering novel mutations in the protein translation of the Fc region. This was particularly high in IGHA1 accounting for 96% of the variants. These allotypes contained a variety of amino acid mutations that can be linked to IgA dimerisation, decreasing or increasing effector function, and could potentially lead to the creation of bispecific antibodies (36–38). Bispecific antibodies are able to target two different epitopes through the union of chains from two different antibodies. This is useful in targeting complex diseases, such as HIV, where antigenic drift presents a significant immunological challenge. Understanding how these variants may

impact effector function is important for informing better design of monoclonal antibodies (mAbs). Monoclonal antibodies are a potential therapeutic avenue for the treatment and prevention of diseases. Many mAbs have been developed for the treatment of autoimmune conditions and cancers which remain challenging to treat with other drugs (39,40). Passive immunisation, the infusion of neutralising antibodies to prevent infection, has also been developed in HIV and RSV (41,42). Mutations in the Fc have been used to increase the half-life of mAbs, create bispecific antibodies, and enhance effector functions; all to improve the efficacy of mAb therapeutics (43–47).

The therapeutic implications for IG gene diversity extend beyond the Fc region, where variable gene diversity has been examined in vaccination responses. It has been established that certain host genetics can lead to favourable responses for vaccination, for example having a duplication IGHV1-69 with favourable alleles leads to better neutralising responses against influenza due to enhanced expression (22). This is especially true for germline targeting vaccines. A prior study has shown that lacking certain IGHV1-2 alleles leads to the inability to produce VRC01-like responses, emphasizing the role of germline variation (48). We have shown, in Chapter 3, that there is variability in the expression of germline genes, beyond sequence variation, and this will be an important consideration as demonstrated in the above example in influenza vaccination. Thus, it is likely true that other potential germline targeting strategies will be dependent on variability in both the expression and genetic sequence of other IGHV genes. In HIV, these strategies aim to target germline genes that are associated with broadly neutralising antibody lineages that bind to various regions on the viral envelope. For example, an immunogen has been designed to target the N332 supersite by engaging PGT121-like precursors which are derived from IGHV4–59/IGLV3–21 (49). This immunogen has yet to be tested in human individuals thus it is unknown whether the efficacy of the immunogen is dependent on certain germline variants of IGHV4–59/IGLV3–21, however, there will likely be some effect of nucleotide or expression variation on the responses.

Overall, this thesis uncovers a great deal of novel genetic diversity in the IG genes of an African population. This remains vital as we strive to make immunogenetic databases more comprehensive and inclusive. The novel diversity detailed in this thesis will be submitted to public databases such as OGRDB and Genbank as these data are published, as with the novel IGHV alleles from Chapter 1 which are already under review by the Inferred Allele Review Committee (IARC) of the AIRR community. These data also contribute to our understanding of what genetic diversity exists in human populations and how this genetic diversity contributes to the functional antibody response. This is important as we move towards

a more informed design of therapeutics such as germline targeting vaccine strategies and aids in beginning to explain demographic differences in vaccine and infection responses.

References

1. Bashirova AA, Zheng W, Akdag M, Augusto DG, Vince N, Dong KL, et al. Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) genes. Genes Immun. 2021 Dec;22(7):327–34.

2. Calonga-Solís V, Malheiros D, Beltrame MH, Vargas L de B, Dourado RM, Issler HC, et al. Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection. Front Immunol. 2019 Jun;10:1161.

3. Engelbrecht E, Rodriguez OL, Shields K, Schultze S, Tieri D, Jana U, et al. Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. Genes Immun. 2024 Aug;25(4):297–306.

4. Gibson WS, Rodriguez OL, Shields K, Silver CA, Dorgham A, Emery M, et al. Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. Genes Immun. 2023 Feb;24(1):21–31.

5. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. Front Immunol. 2020 Sep;11:2136.

6. Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. Nat Commun. 2023 Jul 21;14(1):4419.

7. Scheepers C, Shrestha RK, Lambson BE, Jackson KJL, Wright IA, Naicker D, et al. Ability To Develop Broadly Neutralizing HIV-1 Antibodies Is Not Restricted by the Germline Ig Gene Repertoire. J Immunol. 2015;194(9).

8. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, et al. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. Genes Immun. 2015 Jan;16(1):24–34.

9. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. Am J Hum Genet. 2013 Apr;92(4):530–46.

10. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. 2012;569–604.

11. Lefranc MP. Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of Immunoinformatics. Front Immunol. 2014;5.

12. Watson CT, Glanville J, Marasco WA. The Individual and Population Genetics of Antibody Immunity. Trends Immunol. 2017 Jul;38(7):459–70.

13. Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT, et al. VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Res. 2020 Jan 8;48(D1):D1051–6.

14. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, et al. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev. 2018 Jul;284(1):24–41.

15. Mikocziova I, Greiff V, Sollid LM. Immunoglobulin germline gene variation and its impact on human disease. Genes Immun. 2021;22(4):205–17.

16. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020 Oct;586(7831):741–8.

17. Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K, et al. Genetic substructure and complex demographic history of South African Bantu speakers. Nature Communications. 2021 Apr 7;12(1).

18. Tau T, Wally A, Fanie TP, Ngono GL, Mpoloka SW, Davison S, et al. Genetic variation and population structure of Botswana populations as identified with AmpFLSTR Identifiler short tandem repeat (STR) loci. Sci Rep. 2017 Jul 28;7(1):6768.

19. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 2012;13(5):363–73.

20. Yao R, Zhang C, Yu T, Li N, Hu X, Wang X, et al. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. Molecular Cytogenetics. 2017 Aug 23;10(1).

21. Kumar AA, Loeys B, Van De Beek G, Peeters N, Wuyts W, Van Laer L, et al. varAmpliCNV: analyzing variance of amplicons to detect CNVs in targeted NGS data. Robinson P, editor. Bioinformatics 2022 Nov 28:39(1).

22. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci Rep. 2016 Aug;6(1):20842.

23. Kenter AL, Watson CT, Spille JH. Igh Locus Polymorphism May Dictate Topological Chromatin Conformation and V Gene Usage in the Ig Repertoire. Front Immunol. 2021 May;12:682589.

24. Safonova Y, Shin SB, Kramer L, Reecy J, Watson CT, Smith TPL, et al. Variations in antibody repertoires correlate with vaccine responses. Genome Res. 2022 Apr;32(4):791–804.

25. Medvedovic J, Ebert A, Tagoh H, Tamir IM, Schwickert TA, Novatchkova M, et al. Flexible Long-Range Loops in the VH Gene Region of the Igh Locus Facilitate the Generation of a Diverse Antibody Repertoire. Immunity. 2013 Aug 22;39(2):229–44.

26. Peters JM. How DNA loop extrusion mediated by cohesin enables V(D)J recombination. Curr Opin Cell Biol. 2021 Jun 1;70:75–83.

27. Zhang Q, Wu B, Weng Q, Hu F, Lin Y, Xia C, et al. Regeneration of immunocompetent B lymphopoiesis from pluripotent stem cells guided by transcription factors. Cell Mol

Immunol. 2022 Apr;19(4):492–503.

28. Klatt S, Konthur Z. Secretory signal peptide modification for optimized antibody-fragment expression-secretion in Leishmania tarentolae. Microb Cell Fact. 2012 Jul;11:97.

29. Ling WL, Su CTT, Lua WH, Poh JJ, Ng YL, Wipat A, et al. Signal Peptides and Essential Amino Acids in Antibody Production. Cold Spring Harbor Laboratory. 2020.

30. Jackson KJL, Wang Y, Gaeta BA, Pomat W, Siba P, Rimmer J, et al. Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. Immunogenetics. 2012 Jan;64(1):3–14.

31. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol Baltim Md 1950. 2010 Jun 15;184(12):6986–92.

32. Collins AM, Watson CT. Immunoglobulin Light Chain Gene Rearrangements, Receptor Editing and the Development of a Self-Tolerant Antibody Repertoire. Front Immunol. 2018 Oct;9:2249.

33. Langerak AW, Nadel B, De Torbal A, Wolvers-Tettero ILM, van Gastel-Mol EJ, Verhaaf B, et al. Unraveling the consecutive recombination events in the human IGK locus. J Immunol Baltim Md 1950. 2004 Sep 15;173(6):3878–88.

34. Hehle V, Fraser LD, Tahir R, Kipling D, Wu YC, Lutalo PMK, et al. Immunoglobulin kappa variable region gene selection during early human B cell development in health and systemic lupus erythematosus. Mol Immunol. 2015 Jun;65(2):215–23.

35. Vander Heiden JA, Stathopoulos P, Zhou JQ, Chen L, Gilbert TJ, Bolen CR, et al. Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. J Immunol Baltim Md 1950. 2017 Feb 15;198(4):1460–73.

36. Krah S, Sellmann C, Rhiel L, Schröter C, Dickgiesser S, Beck J, et al. Engineering bispecific antibodies with defined chain pairing. New Biotechnol. 2017 Oct 25;39(Pt B):167–73.

37. Monnet C, Jorieux S, Urbain R, Fournier N, Bouayadi K, De Romeuf C, et al. Selection of IgG Variants with Increased FcRn Binding Using Random and Directed Mutagenesis: Impact on Effector Functions. Front Immunol. 2015 Feb 4;6.

38. Woof JM, Russell MW. Structure and function relationships in IgA. Mucosal Immunol. 2011 Nov;4(6):590–7.

39. Hafeez U, Gan HK, Scott AM. Monoclonal antibodies as immunomodulatory therapy against cancer and autoimmune diseases. Curr Opin Pharmacol. 2018 Aug 1;41:114–21.

41. Paul S, Konig MF, Pardoll DM, Bettegowda C, Papadopoulos N, Wright KM, et al. Cancer therapy with antibodies. Nat Rev Cancer. 2024 Jun;24(6):399–426.

42. Domachowske JB. New and Emerging Passive Immunization Strategies for the Prevention of RSV Infection During Infancy. J Pediatr Infect Dis Soc. 2024 Jul 12;13(Supplement_2):S115–24.

43. Mahomed S, Garrett N, Baxter C, Abdool Karim Q, Abdool Karim SS. Clinical Trials of Broadly Neutralizing Monoclonal Antibodies for Human Immunodeficiency Virus

Prevention: A Review. J Infect Dis. 2021 Feb 1;223(3):370–80.

44. Roopenian DC, Akilesh S. FcRn: the neonatal Fc receptor comes of age. Nat Rev Immunol. 2007 Sep;7(9):715–25.

45. Bournazos S, Chow SK, Abboud N, Casadevall A, Ravetch JV. Human IgG Fc domain engineering enhances antitoxin neutralizing antibody activity. J Clin Invest. 2014 Feb 3;124(2):725–9.

46. Stavenhagen JB, Gorlatov S, Tuaillon N, Rankin CT, Li H, Burke S, et al. Fc optimization of therapeutic antibodies enhances their ability to kill tumor cells in vitro and controls tumor expansion in vivo via low-affinity activating Fcgamma receptors. Cancer Res. 2007 Sep 15;67(18):8882–90.

47. Krah S, Kolmar H, Becker S, Zielonka S. Engineering IgG-Like Bispecific Antibodies—An Overview. Antibodies. 2018 Sep;7(3):28.

48. Lazar GA, Dang W, Karki S, Vafa O, Peng JS, Hyun L, et al. Engineered antibody Fc variants with enhanced effector function. Proc Natl Acad Sci U S A. 2006 Mar 14;103(11):4005–10.

49. Leggat DJ, Cohen KW, Willis JR, Fulp WJ, deCamp AC, Kalyuzhniy O, et al. Vaccination induces HIV broadly neutralizing antibody precursors in humans. Science. 2022 Dec 2;378(6623):eadd6502.

50. Steichen JM, Kulp DW, Tokatlian T, Escolano A, Dosenovic P, Stanfield RL, et al. HIV Vaccine Design to Target Germline Precursors of Glycan-Dependent Broadly Neutralizing Antibodies. Immunity. 2016 Sep 20;45(3):483–96.

Thesis References

A summary of all references used in this thesis

Ahmed AA, Keremane SR, Vielmetter J, Bjorkman PJ. Structural characterization of GASDALIE Fc bound to the activating Fc receptor FcγRIIIa. Journal of Structural Biology. 2016;194(1):78–89.

Alamyar E, Duroux P, Lefranc M-P, Giudicelli V. IMGT® Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In: Christiansen FT, Tait BD, editors. Immunogenetics. Vol. 882. Totowa, NJ: Humana Press; 2012 . p. 569–604. (Methods in Molecular Biology).

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The Generation of Antibody Diversity [Internet]. Nih.gov. Garland Science; 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26860/

Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, Qin K, Fu Y, Huang C-Y, Beigel JH, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Scientific Reports. 2016;6(1):20842.

Bashirova AA, Zheng W, Akdag M, Augusto DG, Vince N, Dong KL, O'hUigin C, Carrington M. Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) genes. Genes & Immunity. 2021;22(7–8):327–334.

Bhiman JN, Anthony C, Doria-Rose NA, Karimanzira O, Schramm CA, Khoza T, Kitchin D, Botha G, Gorman J, Garrett NJ, et al. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. Nature Medicine. 2015;21(11):1332–1336.

Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, Hemmers S, Putintseva EV, Obraztsova AS, Shugay M, et al. Antigen receptor repertoire profiling from RNA-seq data. Nature Biotechnology. 2017;35(10):908–911.

Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, Chudakov DM. MiXCR: software for comprehensive adaptive immunity profiling. Nature Methods. 2015;12(5):380–381.

Bournazos S, Chow S-K, Abboud N, Casadevall A, Ravetch JV. Human IgG Fc domain engineering enhances antitoxin neutralizing antibody activity. Journal of Clinical Investigation. 2014;124(2):725–729.

Boyd SD, Jackson KJL. Predicting Vaccine Responsiveness. Cell Host & Microbe. 2015;17(3):301–307.

Boyd SD, Joshi SA. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Crowe Jr. JE, Boraschi D, Rappuoli R, editors. Microbiology Spectrum. 2014;2(5):2.5.23.

Brusco A, Cinque F, Saviozzi S, Boccazzi C, DeMarchi M, Carbonara AO. The G4 gene is duplicated in 44% of human immunoglobulin heavy chain constant region haplotypes. Human Genetics. 1997;100(1):84–89.

Calmettes G, Weiss JN. A Quantitative Method to Track Protein Translocation between Intracellular Compartments in Real-Time in Live Cells Using Weighted Local Variance Image Analysis. Degtyar VE, editor. PLoS ONE. 2013;8(12):e81988.

Calonga-Solís V, Malheiros D, Beltrame MH, Vargas LDB, Dourado RM, Issler HC, Wassem R, Petzl-Erler ML, Augusto DG. Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection. Frontiers in Immunology. 2019;10:1161.

Carayannopoulos L, Hexham JM, Capra JD. Localization of the binding site for the monocyte immunoglobulin (Ig) A-Fc receptor (CD89) to the domain boundary between Calpha2 and Calpha3 in human IgA1. The Journal of experimental medicine. 1996;183(4):1579–1586.

Chaudhary N, Wesemann DR. Analyzing Immunoglobulin Repertoires. Frontiers in Immunology. 2018;9:462.

Chi X, Li Y, Qiu X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology. 2020;160(3):233–247.

Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies. 2019;8(4):55.

Chung A, Rollman E, Johansson S, Kent S, Stratov I. The Utility of ADCC Responses in HIV Infection. Current HIV Research. 2008;6(6):515–519.

Cohen KW, De Rosa SC, Fulp WJ, deCamp AC, Fiore-Gartland A, Mahoney CR, Furth S, Donahue J, Whaley RE, Ballweber-Fleming L, et al. A first-in-human germline-targeting HIV nanoparticle vaccine induced broad and publicly targeted helper T cell responses. Science Translational Medicine. 2023;15(697):eadf3309.

Collins AM, Wang Y, Singh V, Yu P, Jackson KJ, Sewell WA. The reported germline repertoire of human immunoglobulin kappa chain genes is relatively complete and accurate. Immunogenetics. 2008;60(11):669–676.

Collins AM, Watson CT. Immunoglobulin Light Chain Gene Rearrangements, Receptor Editing and the Development of a Self-Tolerant Antibody Repertoire. Frontiers in Immunology. 2018;9:2249.

Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, Knoetze N, Breden FMW, Christley S, Scott JK, et al. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunological Reviews. 2018;284(1):24–41.

Damelang T, De Taeye SW, Rentenaar R, Roya-Kouchaki K, De Boer E, Derksen NIL, Van Kessel K, Lissenberg-Thunnissen S, Rooijakkers SHM, Jongerius I, et al. The Influence of Human IgG Subclass and Allotype on Complement Activation. The Journal of Immunology. 2023;211(11):1725–1735.

Domachowske JB. New and Emerging Passive Immunization Strategies for the Prevention of RSV Infection During Infancy. Journal of the Pediatric Infectious Diseases Society. 2024;13(Supplement_2):S115–S124.

Doria-Rose NA, Landais E. Coevolution of HIV-1 and broadly neutralizing antibodies. Current Opinion in HIV and AIDS. 2019;14(4):286–293.

Engelbrecht E, Rodriguez OL, Shields K, Schultze S, Tieri D, Jana U, Yaari G, Lees W, Smith ML, Watson CT. Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. 2023 [accessed 2025 Mar 17].

Feng Y, Seija N, Di Noia JM, Martin A. AID in Antibody Diversification: There and Back Again. Trends in Immunology. 2020;41(7):586–600.

Ford M, Haghshenas E, Watson CT, Sahinalp SC. Genotyping and Copy Number Analysis of Immunoglobin Heavy Chain Variable Genes Using Long Reads. iScience. 2020;23(3):100883.

Gibson W, Rodriguez O, Shields K, Silver C, Dorgham A, Emery M, Deikus G, Sebra R, Eichler E, Bashir A, et al. Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. 2022 [accessed 2025 Mar 17].

Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, Sarna VK, Lundin KEA, Clouser C, Vigneault F, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus as revealed by Bayesian haplotyping. 2018 [accessed 2025 Mar 17].

Gutzeit C, Chen K, Cerutti A. The enigmatic function of IgD: some answers at last. European Journal of Immunology. 2018;48(7):1101–1113.

Hafeez U, Gan HK, Scott AM. Monoclonal antibodies as immunomodulatory therapy against cancer and autoimmune diseases. Current Opinion in Pharmacology. 2018;41:114–121.

Hehle V, Fraser LD, Tahir R, Kipling D, Wu Y-C, Lutalo PMK, Cason J, Choong L, D'Cruz DP, Cope AP, et al. Immunoglobulin kappa variable region gene selection during early human B cell development in health and systemic lupus erythematosus. Molecular Immunology. 2015;65(2):215–223.

Hexham JM, White KD, Carayannopoulos LN, Mandecki W, Brisette R, Yang Y-S, Capra JD. A Human Immunoglobulin (Ig)A Cα3 Domain Motif Directs Polymeric Ig Receptor–mediated Secretion. The Journal of Experimental Medicine. 1999;189(4):747–752.

Hoi KH, Ippolito GC. Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire. Genes & Immunity. 2013;14(4):271–276.

Jackson KJL, Wang Y, Gaeta BA, Pomat W, Siba P, Rimmer J, Sewell WA, Collins AM. Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. Immunogenetics. 2012;64(1):3–14.

Janeway CA, Travers P, Walport M, Shlomchik MJ. Immunobiology [Internet]. Nih.gov. Garland Science; 2001. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10757/

Jefferis R, Lefranc M-P. Human immunoglobulin allotypes: Possible implications for immunogenicity. mAbs. 2009;1(4):332–338.

Kawasaki K, Minoshima S, Nakato E, Shibuya K, Shintani A, Schmeits JL, Wang J, Shimizu N. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. Genome Research. 1997;7(3):250–261.

Kenter AL, Watson CT, Spille J-H. Igh Locus Polymorphism May Dictate Topological Chromatin Conformation and V Gene Usage in the Ig Repertoire. Frontiers in Immunology. 2021;12:682589.

Klatt S, Konthur Z. Secretory signal peptide modification for optimized antibody-fragment expression-secretion in Leishmania tarentolae. Microbial Cell Factories. 2012;11(1):97.

Krah S, Kolmar H, Becker S, Zielonka S. Engineering IgG-Like Bispecific Antibodies—An Overview. Antibodies. 2018;7(3):28.

Krah S, Sellmann C, Rhiel L, Schröter C, Dickgiesser S, Beck J, Zielonka S, Toleikis L, Hock B, Kolmar H, et al. Engineering bispecific antibodies with defined chain pairing. New Biotechnology. 2017;39:167–173.

Lazar GA, Dang W, Karki S, Vafa O, Peng JS, Hyun L, Chan C, Chung HS, Eivazi A, Yoder SC, et al. Engineered antibody Fc variants with enhanced effector function. Proceedings of the National Academy of Sciences. 2006;103(11):4005–4010.

Lee PD, Mukherjee S, Edeling MA, Dowd KA, Austin SK, Manhart CJ, Diamond MS, Fremont DH, Pierson TC. The Fc Region of an Antibody Impacts the Neutralization of West Nile Viruses in Different Maturation States. Journal of Virology. 2013;87(24):13729–13740.

Lefranc M-P. Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of Immunoinformatics. Frontiers in Immunology. 2014 [accessed 2025 Mar 17];5.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. IMGT(R), the international ImMunoGeneTics information system(R). Nucleic Acids Research. 2009;37(Database):D1006–D1012.

Lefranc M-P, Lefranc G. Human Gm, Km, and Am Allotypes and Their Molecular Characterization: A Remarkable Demonstration of Polymorphism. In: Christiansen FT, Tait BD, editors. Immunogenetics. Vol. 882. Totowa, NJ: Humana Press; 2012 [accessed 2025 Mar 17]. p. 635–680. (Methods in Molecular Biology™).

Lefranc M-P, Lefranc G. The Immunoglobulin FactsBook. Burlington: Elsevier Science; 2001. (Factsbook).

Leggat DJ, Cohen KW, Willis JR, Fulp WJ, deCamp AC, Kalyuzhniy O, Cottrell CA, Menis S, Finak G, Ballweber-Fleming L, et al. Vaccination induces HIV broadly neutralizing antibody precursors in humans. Science. 2022;378(6623):eadd6502.

Leusen JHW, Nimmerjahn F. The Role of IgG in Immune Responses. In: Nimmerjahn F, editor. Molecular and Cellular Mechanisms of Antibody Activity. New York, NY: Springer New York; 2013 [accessed 2025 Mar 17]. p. 85–112.

Liu J, Wang Y, Xiong E, Hong R, Lu Q, Ohno H, Wang J-Y. Role of the IgM Fc Receptor in Immunity and Tolerance. Frontiers in Immunology. 2019;10:529.

Low TLK, Liu Y-SV, Putnam FW. Structure, Function, and Evolutionary Relationships of Fc Domains of Human Immunoglobulins A, G, M, and E. Science. 1976;191(4225):390–392.

Lu LL, Suscovich TJ, Fortune SM, Alter G. Beyond binding: antibody effector functions in infectious diseases. Nature Reviews Immunology. 2018;18(1):46–61.

Mahomed S, Garrett N, Baxter C, Abdool Karim Q, Abdool Karim SS. Clinical Trials of Broadly Neutralizing Monoclonal Antibodies for Human Immunodeficiency Virus Prevention: A Review. The Journal of Infectious Diseases. 2021;223(3):370–380.

Mascola JR, Haynes BF. HIV -1 neutralizing antibodies: understanding nature's pathways. Immunological Reviews. 2013;254(1):225–244.

Medvedovic J, Ebert A, Tagoh H, Tamir IM, Schwickert TA, Novatchkova M, Sun Q, Huis in 't Veld PJ, Guo C, Yoon HS, et al. Flexible Long-Range Loops in the VH Gene Region of the Igh Locus Facilitate the Generation of a Diverse Antibody Repertoire. Immunity. 2013;39(2):229–244.

Mendelow B, Ramsay M, Chetty N, Stevens W, editors. Molecular Medicine for Clinicians. Wits University Press; 2008 [accessed 2025 Mar 17].

Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, Sollid LM. Polymorphisms in immunoglobulin heavy chain variable genes and their upstream regions. 2020 [accessed 2025 Mar 17].

Mikocziova I, Greiff V, Sollid LM. Immunoglobulin germline gene variation and its impact on human disease. Genes & Immunity. 2021;22(4):205–217.

Mikocziova I, Peres A, Gidoni M, Greiff V, Yaari G, Sollid LM. Alternative splice variants and germline polymorphisms in human immunoglobulin light chain genes. 2021 [accessed 2025 Mar 17].

Mimoto F, Igawa T, Kuramochi T, Katada H, Kadono S, Kamikawa T, Shida-Kawazoe M, Hattori K. Novel asymmetrically engineered antibody Fc variant with superior FcγR binding affinity and specificity compared with afucosylated Fc variant. mAbs. 2013;5(2):229–236.

Molé CM, Béné MC, Montagne PM, Seilles E, Faure GC. Light chains of immunoglobulins in human secretions. Clinica Chimica Acta. 1994;224(2):191–197.

Monnet C, Jorieux S, Urbain R, Fournier N, Bouayadi K, De Romeuf C, Behrens CK, Fontayne A, Mondon P. Selection of IgG Variants with Increased FcRn Binding Using Random and Directed Mutagenesis: Impact on Effector Functions. Frontiers in Immunology. 2015 [accessed 2025 Mar 17];6.

Moore PL, Williamson C. Approaches to the induction of HIV broadly neutralizing antibodies. Current Opinion in HIV and AIDS. 2016;11(6):569–575.

Muramatsu M, Yoshida R, Yokoyama A, Miyamoto H, Kajihara M, Maruyama J, Nao N, Manzoor R, Takada A. Comparison of Antiviral Activity between IgA and IgG Specific to Influenza Virus Hemagglutinin: Increased Potential of IgA for Heterosubtypic Immunity. Kobinger GP, editor. PLoS ONE. 2014;9(1):e85582.

Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, Thörnqvist L, Bürckert J-P, Jackson KJL, Ralph D, et al. Inferred Allelic Variants of Immunoglobulin

Receptor Genes: A System for Their Evaluation, Documentation, and Naming. Frontiers in Immunology. 2019;10:435.

Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT, Boyd SD, Collins AM, Lees W, Yaari G. VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Research. 2020;48(D1):D1051–D1056.

Paul S, Konig MF, Pardoll DM, Bettegowda C, Papadopoulos N, Wright KM, Gabelli SB, Ho M, Van Elsas A, Zhou S. Cancer therapy with antibodies. Nature Reviews Cancer. 2024;24(6):399–426.

Peters J-M. How DNA loop extrusion mediated by cohesin enables V(D)J recombination. Current Opinion in Cell Biology. 2021;70:75–83.

Pleass RJ, Dunlop JI, Anderson CM, Woof JM. Identification of Residues in the CH2/CH3 Domain Interface of IgA Essential for Interaction with the Human Fcα Receptor (FcαR) CD89. Journal of Biological Chemistry. 1999;274(33):23508–23514.

Pollard AJ, Bijker EM. Publisher Correction: A guide to vaccinology: from basic principles to new developments. Nature Reviews Immunology. 2021;21(2):129–129.

Pomarici ND, Cacciato R, Kokot J, Fernández-Quintero ML, Liedl KR. Evolution of the Immunoglobulin Isotypes—Variations of Biophysical Properties among Animal Classes. Biomolecules. 2023;13(5):801.

Prabakaran P, Glanville J, Ippolito GC. Editorial: Next-Generation Sequencing of Human Antibody Repertoires for Exploring B-cell Landscape, Antibody Discovery and Vaccine Development. Frontiers in Immunology. 2020;11:1344.

Richardson SI, Lambson BE, Crowley AR, Bashirova A, Scheepers C, Garrett N, Abdool Karim S, Mkhize NN, Carrington M, Ackerman ME, et al. IgG3 enhances neutralization potency and Fc effector function of an HIV V2-specific broadly neutralizing antibody. Doores KJ, editor. PLOS Pathogens. 2019;15(12):e1008064.

Rispens T, Davies AM, Ooijevaar-de Heer P, Absalah S, Bende O, Sutton BJ, Vidarsson G, Aalberse RC. Dynamics of Inter-heavy Chain Interactions in Human Immunoglobulin G (IgG) Subclasses Studied by Kinetic Fab Arm Exchange. Journal of Biological Chemistry. 2014;289(9):6098–6109.

Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, Deikus G, Auckland K, Eichler EE, Marasco WA, et al. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. 2020 [accessed 2025 Mar 17].

Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, Tieri D, Ke H, Jackson KJL, Boyd SD, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. 2022 [accessed 2025 Mar 17].

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

Roopenian DC, Akilesh S. FcRn: the neonatal Fc receptor comes of age. Nature Reviews Immunology. 2007;7(9):715–725.

Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. Gellert M, Craig N, editors. Microbiology Spectrum. 2014;2(6):2.6.18.

Sadofsky MJ. The RAG proteins in V(D)J recombination: more than just a nuclease. Nucleic Acids Research. 2001;29(7):1399–1409.

Safonova Y, Shin SB, Kramer L, Reecy J, Watson CT, Smith TPL, Pevzner PA. Revealing how variations in antibody repertoires correlate with vaccine responses. 2021 [accessed 2025 Mar 17].

Sangesland M, Ronsard L, Kazer SW, Bals J, Boyoglu-Barnum S, Yousif AS, Barnes R, Feldman J, Quirindongo-Crespo M, McTamney PM, et al. Germline-Encoded Affinity for Cognate Antigen Enables Vaccine Amplification of a Human Broadly Neutralizing Response against Influenza Virus. Immunity. 2019;51(4):735-749.e8.

Scheepers C, Shrestha RK, Lambson BE, Jackson KJL, Wright IA, Naicker D, Goosen M, Berrie L, Ismail A, Garrett N, et al. Ability To Develop Broadly Neutralizing HIV-1 Antibodies Is Not Restricted by the Germline Ig Gene Repertoire. The Journal of Immunology. 2015;194(9):4371–4378.

Schroeder HW, Cavacini L. Structure and function of immunoglobulins. Journal of Allergy and Clinical Immunology. 2010;125(2):S41–S52.

Schwartz GW, Hershberg U. Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics. Physical Biology. 2013;10(3):035005.

Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K, Gunnink H, Chousou-Polydouri N, Delius P, Tollman S, et al. Genetic-substructure and complex demographic history of South African Bantu speakers. 2020 [accessed 2025 Mar 17].

Shamji MH, Valenta R, Jardetzky T, Verhasselt V, Durham SR, Würtzen PA, Van Neerven RJJ. The role of allergen-specific IgE, IgG and IgA in allergic disease. Allergy. 2021;76(12):3627–3641.

Shields RL, Namenuk AK, Hong K, Meng YG, Rae J, Briggs J, Xie D, Lai J, Stadlen A, Li B, et al. High Resolution Mapping of the Binding Site on Human IgG1 for FcγRI, FcγRII, FcγRIII, and FcRn and Design of IgG1 Variants with Improved Binding to the FcγR. Journal of Biological Chemistry. 2001;276(9):6591–6604.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7(1):539.

Stamatatos L, Pancera M, McGuire AT. Germline-targeting immunogens. Immunological Reviews. 2017;275(1):203–216.

Stavenhagen JB, Gorlatov S, Tuaillon N, Rankin CT, Li H, Burke S, Huang L, Johnson S, Bonvini E, Koenig S. Fc Optimization of Therapeutic Antibodies Enhances Their Ability to Kill Tumor Cells *In vitro* and Controls Tumor Expansion *In vivo* via Low-Affinity Activating Fcγ Receptors. Cancer Research. 2007;67(18):8882–8890.

Stavnezer J, Guikema JEJ, Schrader CE. Mechanism and Regulation of Class Switch Recombination. Annual Review of Immunology. 2008;26(1):261–292.

Steichen JM, Kulp DW, Tokatlian T, Escolano A, Dosenovic P, Stanfield RL, McCoy LE, Ozorowski G, Hu X, Kalyuzhniy O, et al. HIV Vaccine Design to Target Germline Precursors of Glycan-Dependent Broadly Neutralizing Antibodies. Immunity.

2016;45(3):483–496.

Talathi SP, Shaikh NN, Pandey SS, Saxena VA, Mamulwar MS, Thakar MR. FcγRIIIa receptor polymorphism influences NK cell mediated ADCC activity against HIV. BMC Infectious Diseases. 2019;19(1):1053.

Tang C, Bagnara D, Chiorazzi N, Scharff MD, MacCarthy T. AID Overlapping and Polη Hotspots Are Key Features of Evolutionary Variation Within the Human Antibody Heavy Chain (IGHV) Genes. Frontiers in Immunology. 2020;11:788.

Tau T, Wally A, Fanie TP, Ngono GL, Mpoloka SW, Davison S, D'Amato ME. Genetic variation and population structure of Botswana populations as identified with AmpFLSTR Identifiler short tandem repeat (STR) loci. Scientific Reports. 2017;7(1):6768.

Tay MZ, Wiehe K, Pollara J. Antibody-Dependent Cellular Phagocytosis in Antiviral Immune Responses. Frontiers in Immunology. 2019;10:332.

Taylor RA, Xiao S, Carias AM, McRaven MD, Thakkar DN, Araínga M, Allen EJ, Rogers KA, Kumarapperuma SC, Gong S, et al. PET/CT targeted tissue sampling reveals virus specific dIgA can alter the distribution and localization of HIV after rectal exposure. Douek DC, editor. PLOS Pathogens. 2021;17(6):e1009632.

Turula H, Wobus CE. The Role of the Polymeric Immunoglobulin Receptor and Secretory Immunoglobulins during Mucosal Infection and Immunity. Viruses. 2018;10(5):237.

Vidarsson G, Dekkers G, Rispens T. IgG Subclasses and Allotypes: From Structure to Effector Functions. Frontiers in Immunology. 2014 [accessed 2025 Mar 17];5.

Wardemann H, Hammersen J, Nussenzweig MC. Human Autoantibody Silencing by Immunoglobulin Light Chains. The Journal of Experimental Medicine. 2004;200(2):191–199.

Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes & Immunity. 2012;13(5):363–373.

Watson CT, Glanville J, Marasco WA. The Individual and Population Genetics of Antibody Immunity. Trends in Immunology. 2017;38(7):459–470.

Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, Wilson RK, Holt RA, Eichler EE, Breden F. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. 2014 [accessed 2025 Mar 17].

Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, et al. Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. The American Journal of Human Genetics. 2013;92(4):530–546.

White KD, Capra JD. Targeting Mucosal Sites by Polymeric Immunoglobulin Receptor-directed Peptides. The Journal of Experimental Medicine. 2002;196(4):551–555.

Woof JM, Russell MW. Structure and function relationships in IgA. Mucosal Immunology. 2011;4(6):590–597.

Yamane-Ohnuki N, Kinoshita S, Inoue-Urakubo M, Kusunoki M, Iida S, Nakano R, Wakitani M, Niwa R, Sakurada M, Uchida K, et al. Establishment of *FUT8* knockout Chinese hamster ovary cells: An ideal host cell line for producing completely defucosylated antibodies with enhanced antibody-dependent cellular cytotoxicity. Biotechnology and Bioengineering. 2004;87(5):614–622.

Zhang Q, Wu B, Weng Q, Hu F, Lin Y, Xia C, Peng H, Wang Y, Liu X, Liu L, et al. Regeneration of immunocompetent B lymphopoiesis from pluripotent stem cells guided by transcription factors. Cellular & Molecular Immunology. 2021;19(4):492–503.

Zhang Y, Zhang X, Ba Z, Liang Z, Dring EW, Hu H, Lou J, Kyritsis N, Zurita J, Shamim MS, et al. The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. Nature. 2019;573(7775):600–604.

Zhou T, Georgiev I, Wu X, Yang Z-Y, Dai K, Finzi A, Do Kwon Y, Scheid JF, Shi W, Xu L, et al. Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01. Science. 2010;329(5993):811–817.

Addenda

Addendum A: Ethics Clearance Certificates

Wits HREC Certificate for this study

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Miss Alaine Marsden et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M220242

NAME: (Principal Investigator)	Miss Alaine Marsden et al
DEPARTMENT:	Virology Centre for HIV & STI's National Institute for Communicable Diseases (NICD)
PROJECT TITLE:	Immunogenetic Diversity in an African Population
DATE CONSIDERED:	25/02/2022
DECISION:	Approved unconditionally
CONDITIONS:	
NOTE:	If contact information regarding student study participants is required, please contact the University Registrar's office <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Dr C. Scheepers
APPROVED BY:	CB Penny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	25/04/2022

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. I **agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **February** and will therefore be due in the month of **February** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature _____ Date _____

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Wits HREC Certificate for CAPRISA cohorts

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Professor P Moore, et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M210892

NAME: (Principal Investigator)	Professor P Moore, et al
DEPARTMENT:	Centre for HIV & Sexually Transmitted Infections National Institute for Communicable Diseases 1 Modderfontein Road Sandringham 2031
PROJECT TITLE:	*Viral set point and clinical disease progression: The role of immunological, genetic and viral factors over the course of disease and during antiretroviral therapy*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Formerly M160971 (expired on 2021/08/04)
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Not applicable
APPROVED BY:	CB Penny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2021/09/07

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **August** and therefore reports and re-certification will be due in the month of **August** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator _____ Date _____

UKZN BREC Certificate for CAPRISA cohorts

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

06 May 2021

Professor S S Abdool Karim
CAPRISA
2nd Floor, DDMRI Building
Nelson R. Mandela School of Medicine
University of KwaZulu-Natal

Dear Professor Abdool Karim

PROTOCOL: CAPRISA 002 Viral set point and clinical progression in HIV-1 subtype C infection: The role of immunological and viral factors during acute and early infection.
REF: E013/04.

RECERTIFICATION APPLICATION APPROVAL NOTICE

Approved:	**30 April 2021**
Expiration of Ethical Approval:	**29 April 2022**

I wish to advise you that your application for recertification received on 08 March 2021 for the above study has been **noted and approved** by a subcommittee of the Biomedical Research Ethics Committee (BREC).The start and end dates of this period are indicated above.

If any modifications or adverse events occur in the project before your next scheduled review, you must submit them to BREC for review. Except in emergency situations, no change to the protocol may be implemented until you have received written BREC approval for the change.

The committee will be notified of the above approval at its next meeting to be held on 08 June 2021.

Yours sincerely



.....
Ms A Marimuthu
(for) Prof D Wassenaar
Chair: Biomedical Research Ethics Committee

Biomedical Research Ethics Committee
Chair: Professor D R Wassenaar
UKZN Research Ethics Office Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Email: BREC@ukzn.ac.za
Website: http://research.ukzn.ac.za/Research-Ethics/Biomedical-Research-Ethics.aspx

Founding Campuses: Edgewood Howard College Medical School Pietermaritzburg Westville

INSPIRING GREATNESS

Addendum B: Supplementary Data from Chapter 1

Supplementary Table 1: Primers for germline NGS.

Primers designed to amplify germline IGHV genes, grouped by target subgroup. The number of genes expected to be amplified by a particular primer pair or grouping is indicated as well as the amplicon size expected to be produced.

Subgroup	Primer Name	Annealing Temperature (C)	Number of Target IGHV Genes	Pooling Ratio	Sequence (5' → 3')	Expected Amplicon Size (bp)	Sequencing Platform
IGHV 1	Multi Forward	65	4	1:2	TGTGCCCTGAGAGCATCACCCA	~570	Illumina MiSeq
	1-2 Reverse				AGCACAGCTGCCTCCTCCCTKG		
	1-3 Reverse				AGCACAGCTGCTTCCCCCCTGG		
	1-18 Reverse				AGTGCAGCTGCCTCCTCCCTGG		
	1-45 Reverse				AGTGAAGCTGCCTCCTCCTTGG		
	1-46 Forward		1	1:2	CCACATCCCTTCTCTACAGAAG		
	1-46 Reverse				GCACAGCTGACTCCTCCCTCAG		
	1-24 Forward		1		CCCAGAGACCATCACACAACAGC		
	1-24 Reverse				CACAGCTGCCTCATTCCTCAG		
	1-69 Forward		2		ATTCCTCCTCTRAAGAAGCCCC		
	1-69 Reverse				GCACAGCTGCCTYCTYCCTCAG		
	1-58 Forward		1		AACCACATCCCTCCGCTAGAG		
	1-58 Reverse				GTACAGCTGCCTCCTTCCTGG		

	Forward	62		1:10	TCACTGCTCCTCACCACAGATG	~860	Pacific Bioscience Sequel 1
	Reverse 1		2		TTTCATGCTSAATCTTTGTCAA		
IGHV 2	Reverse 2		1		TTTCATGCTGAATGTTTGACAA		
	Forward	65		5:10	GACCCTGCAGSTCTSRGAG	~640	
	Reverse 1		7		CTGATCAGTGGCTCCTGAGC		
	Reverse 2		2		TGACTCTGATCAGYGGGTCC		
	Reverse 3		2		TCCTGGGACCTGTCCCGYMC		
	Reverse 4		1		CCTATGCTCAGTGTGTCTTGAG		
	Reverse 5		2		TGCCCTGGAGACAGCCCCGTG		
	Reverse 6		2		TGGACTTGCCCCTGGGGACAG		
	Reverse 7		2		TCTCAGATGCCCCTGCAGCG		
	Reverse 8		2		CCTCCTGTCTTCCTGCAGGG		
IGHV 3	Reverse 9		1		TCGCCCAGTGTTCCTGCAGG		
	Forward 1	64	10	2.5:10	AGACTGGGGSCTCCCTCMTC	~850	
	Forward 2		1		AGACTGGCGCCTCTCTCATC		
IGHV 4	Reverse 1				ATGTGAGGAAGYAGAACCACAS		
	Forward	60		01:10	GATTTATCACCTTCACAGGGTC	~880	
IGHV 5	Reverse		3		GGACACAAACAGCCAGTGTTTC		
	Forward	65		0.5:10	GTTTCCTGTTTAAATGTAGCC	~880	
IGHV 6	Reverse		1		TAGAACTTCTTTGAGGCACACTG		

Supplementary Table 2: IGHV Alleles Detected

This table lists the genes and alleles detected in this study. For representation and analysis, alleles of genes with a designated "D" at the end were treated as alleles of the original gene without "D", e.g. IGHV2-70D and IGHV2-70. This is because these genes are duplications of the original gene. However, these alleles are preserved here so they can be matched with the correct database allele. The genes/alleles are also differentiated by source, either genomic or transcriptomic.

IGHV Gene	Germline Alleles	Transcriptomic Alleles
IGHV1-18	01, p05, 01_T121G	01, 01_A190G
IGHV1-2	02, 06, 04, 05, 02_G301A	02, 04, 06, 06_G34A
IGHV1-24	01	01
IGHV1-3	01, 02, 03, 02_T125G	01, 02
IGHV1-45	02, 03	02
IGHV1-46	01, 03	01, 01_G73A, 03
IGHV1-58	02, p03, 01	02_G61T, 02
IGHV1-69	p30, 10, 14, 04_A163G, 12, p35, p34, 13, p19, p32, p33, 15, p31, p37, 11_A163G, 13_C191T, 15_G48A, 13_G112A, 04_A163G	02, 13_G163A, 02_A18G, 16, 13, 15, 12, 14_G163A, 01, 01_C191T, 04_A244G, 04_A163G, 05
IGHV1-69-2	01, 01_T171C	01, 01_T21C
IGHV1-69D	01	n.d
IGHV2-26	03, 01, p12, 02, p09, 02_G4A, p10	01, 01_C134G, 03
IGHV2-5	05, p11, 01, 02_G60A	01, 02
IGHV2-70	15, p24, 05, p14, 04_C221G, p25, 01, 16, p16, 15_G303A	15, 04, 15_G14A, 16
IGHV2-70D	14	n.d
IGHV3-11	04, 05, 01	04, 01, 05, 06_A240G
IGHV3-13	03, 01, p06	01
IGHV3-15	01	01, 01_G147A
IGHV3-20	04	01, 04
IGHV3-21	n.d	01, 03
IGHV3-23	03, 04, 03_G51T	01, 04
IGHV3-23D	01	n.d
IGHV3-30	p20	02, 02_T201C, 04, 05_G293A, 18

IGHV3-30-3	01	01
IGHV3-30-5	02, 03_G301A	n.d
IGHV3-33	02, 01	01, 08, 06
IGHV3-43	01, p03	01, 02
IGHV3-43D	04, 03	03, 04
IGHV3-48	n.d	01, 03
IGHV3-49	n.d	04, 03
IGHV3-53	02, 01	01, 02
IGHV3-64	n.d	01, 07
IGHV3-64D	n.d	09
IGHV3-66	n.d	01, 02, 01_A193G
IGHV3-7	n.d	01, 03, 05
IGHV3-72	n.d	01, 01_C54T
IGHV3-73	01	01, 02
IGHV3-74	n.d	01, 01_G297A
IGHV3-9	n.d	01, 01_G106A, 03
IGHV4-28	05, p08, 01, p09, 07_T290C, 01_A152C	05
IGHV4-30-2	01, 01_G291A, 01_C21G, 01_G129C	01
IGHV4-30-4	p09	08, 07, 01
IGHV4-31	n.d	03, 03_C198T
IGHV4-34	n.d	01
IGHV4-38-2	p04, 02, 03_A291G, 03_G4C	01, 01_A291G, 03_A291G
IGHV4-39	n.d	01, 07
IGHV4-4	07, p11, 02, 01_C46T, 02_A245C	02, 07, 09
IGHV4-59	01, p19, 08, p15, p18	01, 11, 08, 01_A164G
IGHV4-61	09, 02, 01, 07	02, 09_C163T, 09
IGHV5-10-1	n.d	01
IGHV5-51	p11, 01	01, 03
IGHV6-1	01, p03, 03_C288T, 01_C288T	01

Supplementary Table 3: Identity of Novel IGHV Alleles

A list of the novel alleles identified in this study, along with the dataset that the alleles were identified in, as well as the closest IMGT allele match for each sequence and the identity of the sequence to that allele match.

Novel Allele Designation	Genbank Sequence ID	Accession Number	Closest IMGT Allele Match	Percentage Identity to Closest Match
IGHV1-69*12_G48A	IGHV1-69*novel_2	OR079005	IGHV1-69*12	99.65
IGHV1-3*02_T6C	IGHV1-3*novel_2	OR079011	IGHV1-3*02	99.65
IGHV1-3*03_A58G	IGHV1-3*novel1_2	OR079012	IGHV1-3*03	98.96
IGHV1-69*04_A163G	IGHV1-69*novel.8_1	OR079027	IGHV1-69*04	99.31
IGHV1-69*11_A163G	IGHV1-69*novel.1_1	OR079028	IGHV1-69*11	99.65
IGHV1-69*13_C191T	IGHV1-69*novel.2_1	OR079029	IGHV1-69*13	99.65
IGHV1-69*15_G48A	IGHV1-69*novel.4_1	OR079034	IGHV1-69*15	99.65
IGHV1-69*13_G112A	IGHV1-69*novel.6_1	OR079038	IGHV1-69*13	99.65
IGHV1-69*04_A163G	IGHV1-69*novel.7_1	OR079043	IGHV1-69*04	99.31
IGHV1-18*01_T121G	IGHV1-18*novel_1	OR079045	IGHV1-18*01	99.65
IGHV1-3*03_C309T	IGHV1-3*novel_1	OR079048	IGHV1-3*03	99.65
IGHV1-3*02_T125G	IGHV1-3*novel.1_1	OR079049	IGHV1-3*02	99.65
IGHV1-2*02_G301A	IGHV1-2*novel_1	OR079056	IGHV1-2*02	99.65
IGHV1-69-2*01_T171C	IGHV1-69-2*novel_1	OR079062	IGHV1-69-2*01	99.65
IGHV3-23*03_G51T	IGHV3-23*novel_1	OR079079	IGHV3-23*03	99.65
IGHV2-70*04_C221G	IGHV2-70*novel_1	OR079085	IGHV2-70*04	98.63
IGHV2-70*15_G303A	IGHV2-70*novel_1.1	OR079086	IGHV2-70*15	99.66
IGHV4-30-2*01_G291A	IGHV4-30-2*novel_1	OR079096	IGHV4-30-2*01	99.66
IGHV4-30-2*01_C21G	IGHV4-30-2*novel_1.1	OR079097	IGHV4-30-2*01	98.97
IGHV4-30-2*01_G129C	IGHV4-30-2*novel_1.2	OR079098	IGHV4-30-2*01	99.66
IGHV4-4*01_C46T	IGHV4-4*novel_1	OR079103	IGHV4-4*01	99.65
IGHV4-4*02_A245C	IGHV4-4*novel_1.2	OR079105	IGHV4-4*02	99.65
IGHV2-26*02_G4A	IGHV2-26*novel_1	OR079114	IGHV2-26*02	99.66
IGHV4-28*07_T290C	IGHV4-28*novel_1	OR079131	IGHV4-28*07	99.65
IGHV4-28*01_A152C	IGHV4-28*novel_2	OR079132	IGHV4-28*01	99.65
IGHV6-1*03_C288T	IGHV6-1*novel_4	OR079136	IGHV6-1*03	99.66

IGHV6-1*01_C288T	IGHV6-1*novel_2	OR079137	IGHV6-1*01	99.66
IGHV4-38-2*03_A291G	IGHV4-38-2*novel_1	OR079148	IGHV4-38-2*03	99.65
IGHV4-38-2*03_A291G	IGHV4-38-2*novel_1.1	OR079149	IGHV4-38-2*03	99.65
IGHV4-38-2*03_G4C	IGHV4-38-2*novel_1.2	OR079150	IGHV4-38-2*03	99.65
IGHV3-30-5*03_G301A	IGHV3-33*novel_1	OR079152	IGHV3-30-5*03	99.65
IGHV2-5*02_G60A	IGHV2-5*novel_1	OR079157	IGHV2-5*02	99.31
IGHV1-3*02_C108T	IGHV1-3*02_S4863/1-296	OR078942	IGHV1-3*02	98.96
IGHV1-69*01_C191T	IGHV1-69*01_S0281/1-296	OR078943	IGHV1-69*01	99.65
IGHV1-69*04_A244G	IGHV1-69*04_S5190/1-296	OR078945	IGHV1-69*04	99.65
IGHV1-69*13_G163A	IGHV1-69*13_S4396/1-296	OR078946	IGHV1-69*13	99.65
IGHV1-69-2*01_T21C	IGHV1-69-2*01_S3231/1-296	OR078947	IGHV1-69-2*01	97.22
IGHV2-70*15_G14A	IGHV2-70*15_S4058/1-301	OR078948	IGHV2-70*15	99.66
IGHV3-30*02_T201C	IGHV3-30*02_S9143/1-296	OR078949	IGHV3-30*02	99.65
IGHV3-74*01_G107C	IGHV3-74*01_S7653/1-296	OR078952	IGHV3-74*01	96.18
IGHV1-46*01_G73A	IGHV1-46*01_S4697/1-296	OR078953	IGHV1-46*01	93.75
IGHV1-58*02_G61T	IGHV1-58*02_S3393/1-296	OR078955	IGHV1-58*02	99.65
IGHV1-69*02_A18G	IGHV1-69*02_S6402/1-296	OR078956	IGHV1-69*02	99.31
IGHV2-26*01_C134G	IGHV2-26*01_S1353/1-301	OR078957	IGHV2-26*01	99.66
IGHV3-30*05_G293A	IGHV3-30*05_S4259/1-296	OR078958	IGHV3-30*05	99.65
IGHV3-66*01_A193G	IGHV3-66*01_S4464/1-293	OR078959	IGHV3-66*01	99.65
IGHV3-72*01_C54T	IGHV3-72*01_S2905/1-302	OR078960	IGHV3-72*01	99.66
IGHV4-61*09_C163T	IGHV4-61*09_S5977/1-299	OR078961	IGHV4-61*09	99.31
IGHV3-11*06_A240G	IGHV3-11*06_S8453/1-300	OR078962	IGHV3-11*06	99.65

IGHV3-9*01_G106A	IGHV3-9*01_S4376/1-298	OR078963	IGHV3-9*01	99.65
IGHV4-31*03_C198T	IGHV4-31*03_S8471/1-299	OR078964	IGHV4-31*03	99.66
IGHV4-38-2*01_A291G	IGHV4-38-2*01_S0244/1-294	OR078965	IGHV4-38-2*01	99.65
IGHV4-38-2*03_A291G	IGHV4-38-2*02_S8107/1-296	OR078966	IGHV4-38-2*03	99.65
IGHV1-2*06_G34A	IGHV1-2*06_S5931/1-296	OR078967	IGHV1-2*06	99.65
IGHV1-18*01_A190G	IGHV1-18*01_S4206/1-296	OR078969	IGHV1-18*01	99.65
IGHV3-15*01_G147A	IGHV3-15*01_S2325/1-302	OR078970	IGHV3-15*01	99.66
IGHV1-69*14_G163A	IGHV1-69*14_S5279/1-296	OR078971	IGHV1-69*14	99.65
IGHV3-74*01_G297A	IGHV3-74*01_S4752/1-296	OR078973	IGHV3-74*01	99.65
IGHV4-59*01_A164G	IGHV4-59*01_S6809/1-293	OR078974	IGHV4-59*01	99.65
IGHV1-69*04_A163G	IGHV1-69*04_S4205/1-296	OR078975	IGHV1-69*04	99.65

The transcriptomic data has additional columns for the number of reads, the length distribution (minimum, maximum and mean) and the read error distribution (minimum, maximum and mean) of each inferred novel allele.

Allele	**Number Of Reads**	**Mean Length**	**Max Length**	**Min Length**	**Mean V Error**	**Max V Error**	**Min V Error**
IGHV1-69*04_A163G	4924.0	293.1679528838343	295.0	269.0	3.1756701868399677	52.0	0.0
IGHV1-69*14_G163A	9597.0	293.3591747421069	303.0	266.0	1.1973533395852871	55.0	0.0
IGHV1-58*02_G61T	109.0	292.3302752293578	295.0	283.0	1.4403669724770642	17.0	0.0
IGHV4-31*03_C198T	4035.0	296.5382899628253	312.0	285.0	4.633705080545229	92.0	0.0
IGHV3-74*01_G297A	3755.0	293.8122503328895	304.0	281.0	4.476431424766977	39.0	0.0
IGHV4-59*01_A164G	2798.0	292.57147962830595	307.0	279.0	23.79949964260186	59.0	0.0
IGHV1-69*04_A244G	4995.0	293.5373373733736	295.0	271.0	3.008008008008008	74.0	0.0
IGHV3-66*01_A193G	9723.0	290.838115	328.0	278.0	6.154684768	65.0	0.0

		80787824			075697		
IGHV4-38-2*01_A291G	4011.0	291.75018698578907	302.0	278.0	6.5454998753428075	95.0	0.0
IGHV2-26*01_C134G	4921.0	298.0847388742126	301.0	277.0	2.0170697012802274	46.0	0.0
IGHV1-18*01_A190G	41562.0	293.8743804436745	304.0	266.0	1.951277609354699	71.0	0.0
IGHV1-46*01_G73A	2004.0	294.8797405189621	295.0	275.0	0.6027944111776448	21.0	0.0
IGHV1-2*06_G34A	8130.0	293.91316113161133	295.0	275.0	10.840959409594095	49.0	0.0
IGHV1-3*02_C108T	777.0	294.62033462033463	295.0	274.0	9.752895752895753	59.0	0.0
IGHV3-11*06_A240G	5335.0	296.8140581068416	306.0	285.0	5.8069353327085285	63.0	0.0
IGHV3-72*01_C54T	1346.0	299.464338781575	324.0	288.0	11.407132243684993	51.0	0.0
IGHV3-30*02_T201C	86389.0	293.8465429626457	301.0	266.0	7.498581995392932	86.0	0.0
IGHV1-69*13_G163A	18599.0	293.4259368783268	295.0	266.0	2.385397064358299	56.0	0.0
IGHV1-69*01_C191T	3381.0	293.3617272996155	295.0	268.0	6.1520260278024255	67.0	0.0
IGHV3-9*01_G106A	12313.0	294.6713229919597	306.0	278.0	4.005278973442703	51.0	0.0
IGHV3-15*01_G147A	15696.0	300.00554281345563	310.0	284.0	6.807148318042813	64.0	0.0
IGHV1-69*02_A18G	8883.0	293.47877969154564	295.0	272.0	1.1073961499493414	32.0	0.0
IGHV3-30*05_G293A	14448.0	293.45950996677743	296.0	270.0	6.772563676633444	61.0	0.0
IGHV4-61*09_C163T	9746.0	296.54699363841576	310.0	279.0	3.0115945003078184	67.0	0.0
IGHV3-74*01_G107C	947.0	293.9260823653643	295.0	287.0	1.0823653643083422	23.0	0.0
IGHV4-38-2*03_A291G	10413.0	294.20829732065687	301.0	269.0	7.722078171516373	63.0	0.0
IGHV1-69-2*01_T21C	4319.0	293.0743227598981	295.0	270.0	12.473952303774022	58.0	0.0
IGHV2-70*15_G14A	3422.0	296.9228521332554	300.0	271.0	10.463179427235534	53.0	0.0

Supplementary Table 4: IGHV alleles with matching V regions showing mutations in other regions

This table contains some examples of different IGHV alleles (column 1) which have identical V regions but differ in other parts of the gene. The variation is annotated with the region, if it matches another allele and how many SNVs are contained in that region which differ from the closest reference

Accession Number	Sequence ID	V region allele match	Variation in non V regions
OR079099.1	IGHV4-4*07_3	IGHV4-4*07	V-INTRON: matches 08
OR079100.1	IGHV4-4*07_8	IGHV4-4*07	V-INTRON: matches 08, L-PART1: matches 08
OR079106.1	IGHV4-4*07_2	IGHV4-4*07	V-INTRON: matches 05 with 1 SNV, L-PART1: 1 SNV
OR079108.1	IGHV4-4*07_1	IGHV4-4*07	V-INTRON: matches 05 with 1 SNV, L-PART1: matches 08
OR079116.1	IGHV2-26*02_1	IGHV2-26*02	V-INTRON: 10 SNVs
OR079112.1	IGHV2-26*02_5	IGHV2-26*02	V-INTRON: 1 SNV
OR079123.1	IGHV4-61*07_1	IGHV4-61*07	L-PART1: matches 09
OR079124.1	IGHV4-61*07_1.1	IGHV4-61*07	No Variation
OR079129.1	IGHV4-28*01_3	IGHV4-28*01	No Variation
OR079133.1	IGHV4-28*01_1	IGHV4-28*01	V-INTRON: matches 06 with 1 SNV
OR079142.1	IGHV3-30-3*01_1	IGHV3-30-3*01	L-PART1: 1 SNV
OR079143.1	IGHV3-30-3*01_1.1	IGHV3-30-3*01	No Variation
OR079159.1	IGHV4-59*01_9	IGHV4-59*01	No Variation
OR079161.1	IGHV4-59*01_3	IGHV4-59*01	V-INTRON: matches 11 with 1 SNV

OR078984.1	IGHV1-2*02_8	IGHV1-2*02	No Variation
OR079053.1	IGHV1-2*02_1	IGHV1-2*02	V-INTRON: 1 SNV, L-PART1: 1 SNV
OR079057.1	IGHV1-2*06.1_1	IGHV1-2*06	No Variation
OR079054.1	IGHV1-2*06_1	IGHV1-2*06	V-INTRON: 3 SNVs, L-PART1: 1 SNV
OR078978.1	IGHV1-3*02_24	IGHV1-3*02	No Variation
OR079047.1	IGHV1-3*02_1	IGHV1-3*02	V-INTRON: 2 SNVs
OR079046.1	IGHV1-3*03_1	IGHV1-3*03	No Variation
OR078997.1	IGHV1-3*03_3	IGHV1-3*03	V-INTRON: 1 SNV
OR079016.1	IGHV1-24*01.1_2	IGHV1-24*01	V-INTRON: 2 SNVs
OR079058.1	IGHV1-24*01_1	IGHV1-24*01	V-INTRON: 1 SNVs
OR079059.1	IGHV1-45*02_1	IGHV1-45*02	No Variation
OR078982.1	IGHV1-45*02_9	IGHV1-45*02	L-PART1: 1 SNV
OR078976.1	IGHV1-18*01_35	IGHV1-18*01	1 L-PART1: matches 04 with 1 SNV
OR079044.1	IGHV1-18*01_1	IGHV1-18*01	1 L-PART1: matches 04 with 1 SNV

Supplementary Table 5: IGHV Leader Sequences

This table contains the leader sequences analysed in this study for novel variation. Each leader is annotated by gene with the allele v region matches.

V_region allele matches	Leader Sequence
IGHV1-18*01	ATGGACTGGACCTGGAGCATCCTTTTCTTGGTGGCAGCAGCAACAG
IGHV1-45*02/03	ATGGACTGGACCTGGAGAATCCTCTTCTTGGTGGCAGCAGCCACAG
IGHV1-3*02/03/01	ATGGACTGGACCTGGAGGATCCTCTTTTTGGTGGCAGCAGCCACAG
IGHV1-2*04/02/06/05	ATGGACTGGACCTGGAGGATCCTCTTCTTGGTGGCAGCAGCCACAG
IGHV1-46*01/03	ATGGACTGGACCTGGAGGGTCTTCTGCTTGCTGGCTGTAGCTCCAG
IGHV1-45*02	ATGGACTGGACCTGGAGAATCCTCTTCTTGGTGGCAGCAGTCACAG
IGHV1-69*p35/12/p34/14/12/01/p19/p30/13/15/p32/p33/p31	ATGGACTGGACCTGGAGGTTCCTCTTTGTGGTGGCAGCAGCTACAG
IGHV1-24*01	ATGGACTGCACCTGGAGGATCCTCTTCTTGGTGGCAGCAGCTACAG
IGHV1-58*02/p03	ATGGACTGGATTTGGAGGATCCTCTTCTTGGTGGGAGCAGCGACAG
IGHV1-69-2*01	ATGGACTGCACCTGGAGGATCCTCCTCTTGGTGGCAGCAGCTACAG
IGHV1-58*02	ATGGACTGGATTTGGAGGGTCCTCTTCTTGGTGGGAGCAGCGACAG
IGHV1-58*01	ATGGACTGGATTTGGAGGATCCTCTTCTTGGTGAGAGCAGCGACAG
IGHV1-69*13	ATGGACTGGACCTGGAGGTTCCTCTTTGTGGTGGCAGCTACAG
IGHV3-30*p20/02/01	ATGGAGTTTGGGCTGAGCTGGGTTTTCCTCGTTGCTCTTTTAAGAG
IGHV3-43*01/03/04/p03	ATGGAGTTTGGACTGAGCTGGGTTTTCCTTGTTGCTATTTTAAAAG
IGHV2-70D*14	ATGGACATACTTTGTTCCACGCTCCTGCTACTGACTGTCCCGTCCT
IGHV3-13*03/01/p06	ATGGAGTTGGGGCTGAGCTGGGTTTTCCTTGTTGCTATATTAGAAG
IGHV5-51*p11	ATGGGGTCAACCGCCATCCTCGCCCTCCTCCTGGCTGTTATCCAAG
IGHV5-51*01/p11	ATGGGGTCAACCGCCATCCTCGCCCTCCTCCTGGCTGTT

	CTCCAAG
IGHV3-23*03/04	ATGGAGTTTGGGCTGAGCTGGCTTTTTCTTGTGGCTATTTTAAAAG
IGHV3-73*01/04	ATGGAGTTTGGGCTGAGCTGGGTTTTCCTTGTTGCTATTTTAAAAG
IGHV2-70*15/05/p14/p24/p16	ATGGACATACTTTGTTCCACGCTCCTGCTACTGACTGTCCCGTCCT
IGHV2-70*p24/p25	ATGGACATGCTTTGTTCCATGCTCCTGCTACTGACTGTCCCGTCCT
IGHV2-70*p25	ATGGACATGCTTTGTTCCATGCTCCTGCTACTGACTGTCCCGTCTT
IGHV3-11*04/05/01	ATGGAGTTTGGGCTGAGCTGGGTTTTCCTTGTTGCTATTATAAAAG
IGHV4-30-2*01	ATGAAACACCTGTGGTTCTTCCTCCTGCTGGTGGCAGCTCCCAGAT
IGHV4-4*07	ATGAAACACCTGTGGTTCTTCCTCCTGCTGGTGGCAGCTCCCAGAT
IGHV4-4*07/p11/02	ATGAAACACCTGTGGTTCTTCCTCCTCCTGGTGGCAGCTCCCAGAT
IGHV4-4*07	ATGAAGCACCTGTGGTTCTTCCTCCTGCTGGTGGCAGCTCCCAGAT
IGHV4-4*02	ATGAAACATCTGTGGTTCTTCCTCCTCCTGGTGGCAGCTCCCAGAT
IGHV2-26*03/01	ATGGACACACTTTGCTACACACTCCTGCTGCTGACCACCCCTTCCT
IGHV2-26*p12/02/p09/p10	ATGGACACGCTTTGCTACACACTCCTGCTGCTGACCACCCCTTCCT
IGHV2-26*02	ATGGACATGCTTTGTTCCATGCTCCTGCTACTGACTGTCCCGTCCT
IGHV4-30-4*p09	ATGAAACACCTGTGGTTCTTCCTCCTGCTGGTGGCAGCTCCCAGAT
IGHV4-61*07	ATGAAACATCTGTGGTTCTTCCTTCTCCTGGTGGCAGCTCCCAGAT
IGHV4-61*02	ATGAAACATCTGTGGTTCTTCCTCCTCCTGGTGGCAGCTCCCAGAT
IGHV4-61*01	ATGAAACACCTGTGGTTCTTCCTCCTCCTGGTGGCAGCTCCCAGAT
IGHV4-61*07	ATGAAGCACCTGTGGTTCTTCCTCCTGCTGGTGGCGGCTCCCAGAT
IGHV4-28*05/p08/01/p09	ATGAAACACCTGTGGTTCTTCCTCCTGCTGGTGGCAGCTCCCAGAT
IGHV6-1*01/p03	ATGTCTGTCTCCTTCCTCATCTTCCTGCCCGTGCTGGGCCTCCCAT
IGHV3-53*02	ATGGAGTTTTGGCTGAGCTGGGTTTTCCTTGTTGCTATTT

	CAAAAG
IGHV3-53*01	ATGGAGTTTTGGCTGAGCTGGGTTTTCCTTGTTGCTATTTTAAAAG
IGHV3-23D*01	ATGGAGTTTGGGCTGAGCTGGCTTTTTCTTGTGGCTATTTTAAAAG
IGHV3-30-3*01	ATGGAGTTTGGGCTGAGGTGGGTTTTCCTCGTTGCTCTTTTAAGAG
IGHV4-38-2*p04/02	ATGAAGCACCTGTGGTTTTTCCTCCTGCTGGTGGCAGCTCCCAGAT
IGHV2-5*05/p11/01	ATGGACACACTTTGCTCCACGCTCCTGCTGCTGACCATCCCTTCAT
IGHV3-15*01	ATGGAGTTTGGGCTGAGCTGGATTTTCCTTGCTGCTATTTTAAAAG
IGHV4-59*01/p15/p18/08	ATGAAACATCTGTGGTTCTTCCTTCTCCTGGTGGCAGCTCCCAGAT
IGHV4-59*p19|IGHV4-59_2	ATGAAACACCTGTGGTTCTTCCTCCTCCTGGTGGCAGCTCCCAGAT

Supplementary Table 6: Alleles of duplicated IGHV genes from genomic data

The participants that contain duplications detected in the genomic sequencing of this study are listed here, with the alleles that contribute to the detection of a duplication. Of note, where there are alleles listed twice, this is because while the V-region designation for the alleles are identical the sequences differ in other regions of the gene.

CAP Participant	IGHV5-51	IGHV2-70	IGHV4-4	IGHV2-26	IGHV6-1	IGHV4-38-2	IGHV4-59	IGHV1-69	IGHV1-69D	IGHV1-3	IGHV1-46
CAP121	none	none	none	['01']	['01']	none	none	['10', '12', '13', '11_A163G']	none	['02_T125G']	none
CAP257	['01', 'p11']	none	none	none	none	none	none	['p30', 'p35', 'p19']	none	['02']	none
CAP117	['01', 'p11']	['p25']	['07', 'p11']	['01']	none	none	['01', '01', '08']	none	none	none	none
CAP252	['p11']	none	['07']	none	none	['02']	['01']	['14', 'p35', 'p19']	none	none	['01', '01', '03']
CAP8	['p11', '01', 'p11']	['15', 'p24']	['07']	['p12']	['01']	['p04', '02', 'p04']	['01']	none	none	['01', '02', '03', '02_T125G', '02']	none
CAP290	none	none	none	none	none	none	none	none	none	none	none
CAP312	none	none	none	none	['03_C288T', '01_C288T']	none	none	none	none	none	none
CAP248	none	none	none	none	none	['p04']	none	none	none	none	none

CAP221	['p11', '01']	none	none	none	none	none	none	none	none	none	none
CAP85	['p11']	['15']	none	['p12']	none	none	none	none	none	['02', '03', '02_T125G', '02']	none
CAP306	['p11', 'p11']	none	none	none	none	none	none	none	none	none	none
CAP301	['p11']	none	none	none	none	none	none	none	none	none	none
CAP331	['p11', '01']	none	none	none	none	none	none	['12', '13_G163A']	['01']	none	['01']
CAP65	['p11']	none	none	none	['03_C288T', '01_C288T']	none	none	['p30', '10', '14', '04_A163G']	none	['02', '03']	none
CAP109	['p11', 'p11']	['15']	[02_A245C']	none	['01']	['p04', '02']	none	['12', '13']	none	none	none
CAP148	['p11']	none	['07', '07']	none	none	['02', 'p04']	['p19']	none	none	none	none
CAP105	['01', 'p11']	['15']	none	['01']	['01']	['02']	none	none	none	none	none
CAP120	['01']	['05', 'p14', '01']	none	['01']	['01']	none	none	['p34', '01_T21C']	['01']	none	none
CAP64	['01']	none	none	none	['01']	none	none	['12', 'p35']	none	['01', '02', '02_T125G']	['01']

CAP45	['01', 'p11']	none	none	none	none	none	none	none	none	none	none
CAP225	['01', 'p11']	['p14', 'p25']	none	['01']	['p03']	none	none	['12', 'p32', 'p33']	none	['01']	['01']
CAP88	['01', 'p11']	['p24']	['07']	none	['01']	['p04', '02', 'p04']	['01']	none	none	['01', '02']	none
CAP108	['01']	['p24', '05']	['02', '07']	['01']	['01']	['p04', '02']	['01']	none	none	none	none
CAP310	['01']	['15']	['07']	['01']	['01']	['p04']	['p15']	none	none	none	none
CAP244	['p11']	none	['07', '07']	none	['01']	['p04']	['01']	['12', '13']	none	none	['01', '03']
CAP287	['p11']	none	['07']	none	none	['p04', '02', 'p04']	['01']	none	none	['01']	none
CAP268	['p11', 'p11']	['15']	none	['01', '02']	['01', 'p03']	none	none	none	none	['01', '02']	none
CAP286	none	['15']	none	['03', '01']	none	none	none	none	none	none	none
CAP280	none	['15']	none	['01']	none	none	none	none	none	none	none
CAP161	none	['15']	['07', 'p11']	['01']	['01', 'p03']	['02','03_A291G', '03_G4C']	none	['10', '15', '11_A163G']	none	none	['01']
CAP119	none	['15', 'p14', '15_G14A']	none	['01']	none	none	none	['12', 'p35', 'p37']	none	none	none

CAP145	none	['15']	['02']	['01', 'p10']	['p03']	['p04', 'p04']	['01', 'p15']	none	none	none	none
CAP146	none	['15']	['p11']	['02']	['01']	none	['p15', 'p18']	['14', '15', 'p31']	none	none	['01']
CAP3	['p11', '01', 'p11']	['15', 'p24', '05', '16', '15_G14A']	['07']	['01', '02', '02']	['01', 'p03', '03_C288T', '01_C288T']	['p04', '02']	['p19', 'p15']	['10', '12', 'p35', 'p34', '13', '15_G48A', '13_G112A', '04_A163G']	['01']	['01', '02', '02_T125G', '03']	['01']
CAP103	none	['p24', '05']	none	['01', '02', 'p09', '02_G4A']	['01', 'p03']	none	none	['p30', 'p34']	none	none	none
CAP387	none	['p24', '05', '16', '15_G14A']	none	['01', '02', '02']	none	none	none	none	none	none	none
CAP254	none	['p14']	none	['01']	none	none	none	none	none	none	none
CAP228	none	none	['07']	none	none	['p04', 'p04']	['01']	none	none	['01']	none
CAP165	none	none	['07', '02', '02_A245C]	none	none	['02']	['p18']	none	none	none	none
CAP330	none	none	none	none	none	['02']	['p19']	none	none	none	none

CAP118	none	none	['07', '01_C46T']	none	none	['p04', '02', '01_C288T']	['01', '01', '08']	['10', '12', '13_C191T']	none	['01']	['01']
CAP172	none	none	none	none	['01']	none	none	['p34', '15', 'p31']	none	none	['01']
CAP103.2	none	none	none	none	['01', 'p03']	none	none	none	none	none	none
CAP255	none	none	none	none	['01']	none	none	none	none	none	none
CAP335	none	none	none	none	['p03']	none	none	none	none	['01']	none
CAP110	none	none	none	none	none	none	none	none	none	none	none
CAP250	none	none	none	none	none	none	none	none	none	none	none
CAP137	none	none	none	none	none	['02']	none	['10', '12']	none	none	none
CAP80	none	none	none	none	none	none	none	none	none	none	none
CAP383	none	none	none	none	none	none	none	none	none	['02']	none
CAP378	none	none	none	none	none	none	none	none	none	['01', '03']	none
CAP365	none	none	none	none	none	none	none	none	none	['01']	none
CAP363	none	none	none	none	none	none	none	['p35', '11_A163G', 'p34']	none	['01']	none
CAP349	none	none	none	none	none	none	none	none	none	['02']	none

CAP346	none	none	none	none	none	none	none	none	none	['01', '02']	none
CAP342	none	none	none	none	none	none	none	['p35', '13']	none	['01']	none
CAP326	none	none	none	none	none	none	none	none	none	['01', '02_T125G']	none
CAP314	none	none	none	none	none	none	none	none	none	['02']	none
CAP62	none	none	none	none	none	none	none	none	none	['02', '02_T125G']	none
CAP295	none	none	none	none	none	none	none	none	none	['01', '02']	none
CAP229	none	none	none	none	none	none	none	none	none	none	none
CAP288	none	none	none	none	none	none	none	none	none	['01']	none
CAP220	none	none	none	none	none	none	none	none	none	['02_T125G']	none
CAP341	none	none	none	none	none	none	none	['10', 'p35', '15_G48A']	none	none	['01']
CAP115	none	none	none	none	none	none	none	['10', '04_A163G', 'p35']	none	['01']	none
CAP177	none	none	none	none	none	none	none	['12']	none	none	none
CAP147	none	none	none	none	none	none	none	['p35']	none	none	none

CAP357	none	none	none	none	none	none	none	none	none	['01']	none
CAP222	none	none	none	none	none	none	none	none	none	['03']	none

Supplementary Table 7: RabHit output table

This table is large and complex, for ease of reading and accessibility it can be viewed through this link: Supplementary Table 7

Addendum C: Plagiarism Report

Document Viewer

Turnitin Originality Report

Processed on: 12-Mar-2025 2:25 PM SAST
ID: 2612567334
Word Count: 24476
Submitted: 1

PhD_Thesis_Assembly.docx By Joshua Marsden

Similarity Index	Similarity by Source	
4%	Internet Sources:	4%
	Publications:	5%
	Student Papers:	2%

include quoted | include bibliography | excluding matches < 1% | mode: quickview (classic) report | print | download

1% match ()
Alaine A. Marsden, Martin Corcoran, Gunilla Karlsson Hedestam, Nigel Garrett et al. "Novel polymorphic and copy number diversity in the antibody IGH locus of South African individuals", Immunogenetics

1% match (Internet from 06-Apr-2021)
http://wiredspace.wits.ac.za

1% match (Internet from 25-Aug-2024)
https://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-024-01073-y

1% match (student papers from 26-Oct-2021)
Submitted to Universidad de Burgos UBUCEV on 2021-10-26

1% match ()
Haynes, Barton F, Wiehe, Kevin et al. "Strategies for HIV-1 vaccines that induce broadly neutralizing antibodies", Springer Nature, 2022